Robust Speech/Non-Speech Discrimination Based on Pitch Estimation for Mobile Robots

François Grondin and François Michaud

***Abstract*— To be used on a mobile robot, speech/non-speech discrimination must be robust to environmental noise and to the position of the interlocutor, without necessarily having to satisfy low-latency requirements. To address these conditions, this paper presents a speech/non-speech discrimination approach based on pitch estimation. Pitch features are robust to noise and reverberation, and can be estimated over a few seconds. Results suggest that our approach is more robust compared to the use of Mel-Frequency Cepstrum Coefficients with Gaussian Mixture Models (MFCC-GMM) under high reverberation levels and additive noise (with an accuracy above 98% with a latency of 2.21 sec), which makes it ideal for mobile robot applications. The approach is also validated on a mobile robot equipped with a 8-microphone array, using speech/non-speech discrimination based on pitch estimation as a post-processing module of a localization, tracking and separation system.**

I. Introduction

Real world environments are filled with a wide variety of sounds that can be categorized as either speech or non-speech sources (e.g., music, phone ring, door closing, fan noise). Robots can certainly benefit from analyzing these sounds to acquire various types of information about the world, such as: 1) what are the sound sources and where are they coming from; 2) what kinds of sources are making the sounds; and 3) what information can be extracted from these sound sources. For Type 1, localization, tracking and separation of sound sources are typical tasks performed by robot audition systems such as ManyEars [1] or HARK [2]. For Type 3, speech recognition can be performed to recognize the words pronounced by a speaker [3], [4], [5], while speaker identification can provide the identity of the speaker regardless of the words pronounced [6]. For non-speech sources, music recognition and beat tracking have been demonstrated [7], [8], as recognition of daily sounds [9], [10]. Sasaki et al. [10] also propose to identify a sound from a pre-trained database composed of daily sounds and speech. However, their approach requires an offline database of sounds for training, and is generally sensitive to the environment [9], [10], [11]. Before attempting to extract information from the sound sources, it may be more appropriate to discriminate between speech and non-speech (Type 2) to then select the proper recognition algorithm to apply on the audio streams to process, and incidentally improve robustness, processing time and adaptivity to the environment of the artificial audition system.

Voice Activity Detection (VAD) is a type of speech/non-speech discrimination algorithm performed on signals captured by close-talking devices (with no or little reverberation) to initiate speech coding or speech recognition, with minimal latency. VAD can be done using trained support vector machine with subband signal-to-noise ratios (SNRs) extraction through denoising and contextual subband feature extraction [12]. Discrimination may also be performed using the integrated bispectrum and the average likelihood ratio [13], as well as with fuzzy logic with denoised subband SNRs and zero crossing rates [14], and multiscale spectro-temporal modulations [15]. Applying these techniques to speech/non-speech discrimination on robots is difficult because of the presence of noise and reverberation in the signals, and the lack of having trained models that include information about environmental noise. Brueckmann et al. [16] and Heck et al. [17] present approaches that uses Mel-Frequency Cepstral Coefficients (MFCC) and pretrained models of daily sound events to discriminate from speech. However, MFCC features are sensitive to noise [18], [19], and speech/non-speech discrimination is limited to the models of daily sounds available.

For human-robot interaction, low-latency is a soft requirement for speech/non-speech discrimination because a sound source can be tracked over many seconds and the classification may be performed after a few seconds to determine whether or not the source originates from a human talking or some daily sound events (hand claps, phone ringing, door slam, etc.). Speech is made of voiced and unvoiced segments, and a large observation window guarantees that some voiced segments are captured. In these conditions, pitch, present sporadically in speech, becomes a suitable feature for speech/non-speech discrimination as it is robust to channel modulation and reverberation. Also, the fundamental frequency and the harmonics of speech usually have high SNRs, and are therefore robust to background noise. The presence of pitch features makes it therefore possible to discriminate easily between speech and non-speech signals, without requiring pre-identified models of daily sounds. Nielsen et al. [20] present a method that uses pitch and an additional reliability feature to classify sounds as speech, music or noise. Although this method is robust to channel distortion, it is not suitable for noisy environments because the reliability feature used, as opposed to pitch, is sensitive to background noise.

This paper describes a new speech/non-speech discrimi-

This work was supported by the Fonds de recherche du Québec - Nature et technologies (FRQNT).

F. Grondin and F. Michaud are with the Department of Electrical Engineering and Computer Engineering, Interdisciplinary Institute for Technological Innovation (3IT), 3000 boul. de l'Université, Sherbrooke, Québec (Canada) J1K 0A5, {Francois.Grondin2,Francois.Michaud}@USherbrooke.ca

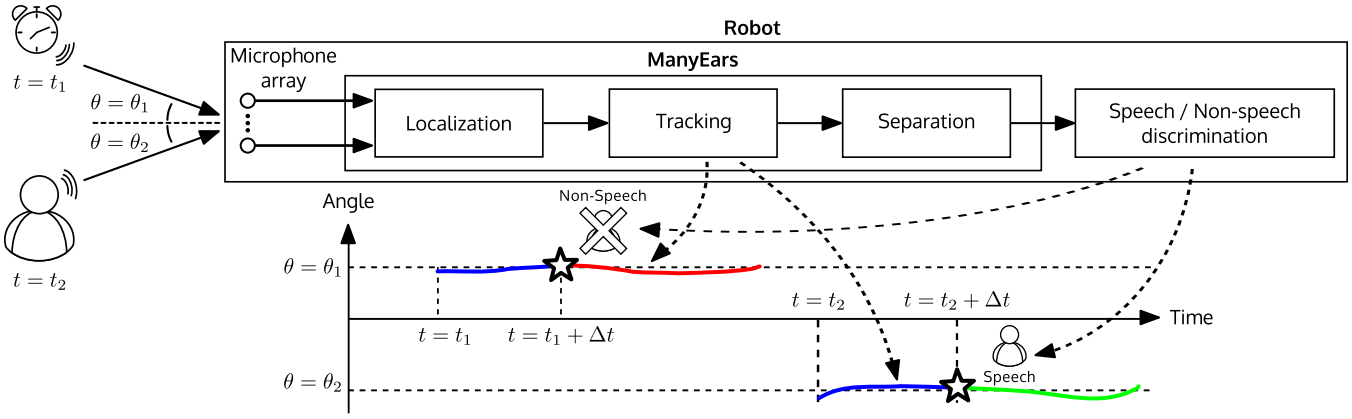


Fig. 1: ManyEars with a speech/non-speech discrimination module

nation approach based on pitch estimation, which only uses pitch features, as opposed to Nielsen et al. [20] which also make use of a reliability feature. The paper is organized as follows. Section II presents an overview of the system in which speech/non-speech discrimination is used for robot audition. Section III demonstrates the pitch features proposed, and Section IV describes how these features are used to perform speech/non-speech discrimination. Section V compares using simulations the proposed method with speech/non-speech discrimination using MFCC features, and demonstrates the performance of the proposed method on a mobile robot platform.

II. System Overview

Figure 1 provides an overview of the system. It is based on ManyEars [1], which uses a 8-microphone array to perform sound source localization, tracking and separation on mobile robots. Sound source localization is done using General Cross Correlation with Phase Transform Weighting (GCC-PHAT) and provides potential positions in space for the active sound sources. Tracking is implemented using particle filters to follow in space one or many active sources as they move over time. The Separation module then reduces the interference from competing sources and enhances the sound stream of each individual source using Geometric Source Separation (GSS). When only one source is active, GSS is then similar to a delay-and-sum beamformer, which reduces reverberation and improves the SNR of the active sound source. Then, the objective of the speech/non-speech discrimination module is to classify the complete tracked source segment as speech or non-speech.

To provide as an example, Figure 1 also illustrates a simple case scenario involving an alarm clock located at the angle $\theta = \theta_1$ that starts to emit sounds at time $t = t_1$. The source is tracked and after a time delay (i.e., latency) of Δt seconds, it is classified as non-speech by the system. An interlocutor then starts talking at $t = t_2$ seconds and is classified as speech after the same time delay. The speech/non-speech information may then be used by the robot to perform automatic speech recognition only on speech sources.

III. Pitch Features

The pitch p is normally given in Hz but can also be expressed as a time delay in seconds $(1/p)$ or in samples (F_s/p) – the latter is used to explain our approach. To derive pitch features, a rectangular window of N samples is applied on the discrete-time input signal $x[n]$ from an unknown source. Short-Time Fourier Transform (STFT) coefficients $X_i[k]$ of the signal at each frame i and bin k are calculated using (1). The variables ΔN and j stand for the hop size and the imaginary number $\sqrt{-1}$, respectively.

$$X_i[k] = \sum_{n=0}^{N-1} x[(i-1)\Delta N + n + 1] \exp\left(-j\frac{2\pi kn}{N}\right) \quad (1)$$

The autocorrelation $R_i[n]$ is computed according to (2), where the operator * stands for the complex conjugate.

$$R_i[n] = \frac{1}{N}\sum_{k=0}^{N-1} X_i[k]X_i[k]^* \exp\left(j\frac{2\pi kn}{N}\right) \quad (2)$$

Empirically, we set q to keep peaks that dominate the three neighboring bins as potential pitch delay candidates as expressed by (3), and the index of the peak with the greatest value is selected according to (4).

$$\hat{R}_i[n] = \begin{cases} R_i[n] & R_i[n] > R_i[n+q], q = \pm 3, \pm 2, \pm 1 \\ 0 & otherwise \end{cases} \quad (3)$$

$$\hat{\tau}_i = \arg\max_n \left(\hat{R}_i[n]\right) \quad (4)$$

The variable $\hat{\tau}_i$ is an approximation of the exact maximum delay τ_i of the autocorrelation caused by discrete sampling. In fact, $\hat{\tau}_i$ is a random variable with a uniform distribution between $\tau_i - 0.5$ and $\tau_i + 0.5$. This error is negligible when the time delay is estimated, but influences significantly the estimation of the time delay difference in time $(\Delta\tau_i)$. A new delay $\widetilde{\tau}_i$ is obtained from quadratic interpolation as expressed by (5), which reduces the error introduced by discretization of the time delay [21].

$$\widetilde{\tau}_i = \hat{\tau}_i + \left(\frac{R_i[\hat{\tau}_i - 1] - R_i[\hat{\tau}_i + 1]}{2R_i[\hat{\tau}_i - 1] - 4R_i[\hat{\tau}_i] + 2R_i[\hat{\tau}_i + 1]} \right) \quad (5)$$

IV. Speech/Non-Speech Discrimination Based on Pitch Estimation and MFCC features

Our approach is based on three realistic working hypotheses related to pitch within speech signals [22]:

- H_1: The pitch of human speech for an adult (man or female) usually lies between 80 Hz and 500 Hz.
- H_2: The pitch of human speech varies slowly in time.
- H_3: Pitch is observed in voiced segments, and the ratio of the duration of these segments over the overall signal duration is significant.

The strategy is to identify speech frames that are good voiced frame candidates according to hypotheses H_1 and H_2. The ratio of the number of voiced frames over all frames is then computed over a finite time window. If this ratio is high enough to satisfy hypothesis H_3, then the signal within the finite time window is classified as speech. This classification method introduces some latency since a representative amount of frames must be gathered before a decision is made. However, as explained previously, the objective is to classify a whole sound segment obtained from the localization, tracking and separation system, and not to perform a frame-by-frame VAD.

A frame is considered to be a voiced frame candidate if the delay lies within the pitch human range (defined by the parameters τ_{min} and τ_{max}) given by hypothesis H_1. The time delay difference $\Delta\widetilde{\tau}_i$ is also investigated to assess hypothesis H_2, and is obtained by (6).

$$\Delta\widetilde{\tau}_i = \widetilde{\tau}_i - \widetilde{\tau}_{i-1} \quad (6)$$

The instantaneous delay difference $\Delta\widetilde{\tau}_i$ gives some indications about the pitch dynamics, but is often insufficient to identify long-term variations in the delay. In fact, in speech signals, the delay usually increases or decreases over a large period of time, while in non-speech signals $\Delta\widetilde{\tau}_i$ may oscillate between positive and negative values without a significant net increase or decrease. To capture long-term variations, a smoothed time delay variable ϕ_i is introduced in (7). When the delay difference magnitude is smaller than a fixed threshold ($\Delta\tau_{max}$), ϕ_i is initialized to $\Delta\widetilde{\tau}_i$ and is recursively updated according to the rate set by α. The magnitude of the delay difference is discarded when it is above $\Delta\tau_{max}$, as it usually indicates a jump between random delay values.

$$\phi_i = \begin{cases} (1-\alpha)\phi_{i-1} + \alpha\Delta\widetilde{\tau}_i & |\Delta\widetilde{\tau}_i| < \Delta\tau_{max} \\ & |\Delta\widetilde{\tau}_{i-1}| < \Delta\tau_{max} \\ \Delta\widetilde{\tau}_i & |\Delta\widetilde{\tau}_i| < \Delta\tau_{max} \\ & |\Delta\widetilde{\tau}_{i-1}| \geq \Delta\tau_{max} \\ 0 & |\Delta\widetilde{\tau}_i| \geq \Delta\tau_{max} \end{cases} \quad (7)$$

When $\widetilde{\tau}_i$ is within speech range and ϕ_i is large enough to match speech delay variation (greater than the fixed threshold ϕ_{min}), the frame i is considered to be a speech voiced frame. These conditions are summarized by (8), where $v_i = 1$ stands for a voiced frame and $v_i = 0$ an unvoiced frame.

$$v_i = \begin{cases} 1 & \tau_{min} \leq \widetilde{\tau}_i \leq \tau_{max}, |\phi_i| > \phi_{min} \\ 0 & otherwise \end{cases} \quad (8)$$

The ratio r of voiced frames in a window size of M frames is defined by (9).

$$r = \frac{1}{M} \sum_{i=0}^{M-1} v_i \quad (9)$$

An utterance is classified as speech when this ratio exceeds a fixed threshold (r_0, between 0 and 1) and considered as non-speech sound otherwise, as expressed by (10).

$$d = \begin{cases} \text{speech} & r > r_0 \\ \text{non speech} & r \leq r_0 \end{cases} \quad (10)$$

V. Results

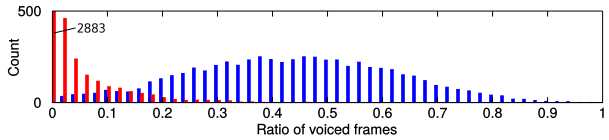
To consider a wide range of noise and reverberation conditions, tests of our speech/non-speech discrimination module were first done in simulation, to then validate the entire system on a mobile robot. Table I presents the parameters used with our approach. These parameters were fixed empirically as follows. The sample rate F_s matches the sample rate used by ManyEars. Parameter N is set to have an analysis window of 85 msec that captures long pitch time periods of male speakers. The hopsize ΔN is selected to ensure a significant overlap between frames. The parameters τ_{min} and τ_{max} are selected to define the pitch range between 80 Hz and 500 Hz. A large value $\Delta\tau_{max}$ is used to detect random jumps in $\Delta\widetilde{\tau}_i$, and we observed pitch variations in speech and non-speech signals to set α and ϕ_{min}.

TABLE I: Parameters of our pitch speech/non-speech discrimination module

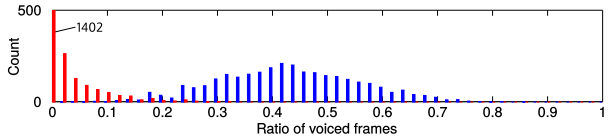
Parameter	Value
F_s	48000
N	4096
ΔN	512
τ_{min}	96
τ_{max}	600
$\Delta\tau_{max}$	20
α	0.3
ϕ_{min}	0.5

A. Simulation

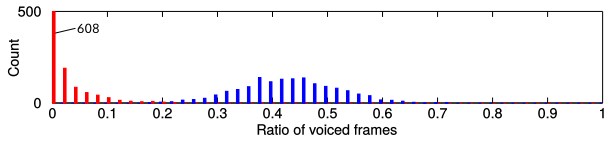
The reference signal used for the simulated experiments consists of 56 minutes of male and female speech and 19 minutes of daily event sounds (hand clap, footsteps, phone ringing, alarm clock, door slam, etc.). Half the data in speech and daily events are chosen for training, and the other half is used for testing. MFCC-GMM training is performed using



(a) $M = 50$ (latency ≈ 0.61 sec)



(b) $M = 100$ (latency ≈ 1.14 sec)



(c) $M = 200$ (latency ≈ 2.21 sec)

Fig. 2: Histograms of the ratio of voiced (blue) and unvoiced (red) frames according to M

the Expectation Maximization (EM) algorithm [23] applied on clean data, while no explicit training is required for the proposed classification system based on pitch (the parameter r_0 is chosen empirically to be 0.12). Testing is then performed on noisy data, which are corrupted by reverberation and noise. Room impulse responses are generated using the Allen and Berkley image method to simulate reverberation [24]. Three types of noise are added: white noise, fan noise and burst noise. Burst noise is made of bursts of one second of white gaussian noise, spaced by silence periods of one second. Figure 2 illustrates the distribution of the ratio of voiced frames (r) for three window sizes (M). M needs to be large enough to ensure good discrimination between speech and non-speech, but small enough to keep latency under an acceptable time interval. As shown, when M increases, the discrimination between the ratios of voiced frames for speech and non-speech sources increases. Setting $M = 200$ provides good discrimination and an acceptable latency slightly above 2 sec.

We also compared the proposed method with the Mel Frequency Cepstrum Coefficients features (MFCC) [17], [25]. Heck et al. [17] use frames of 1024 samples separated by a hop size of 512 samples. The power of the spectrum is multiplied by a filterbank of 24 triangular filters mapped on a Mel scale. The log value of the power in each filter is computed, and then unitary discrete cosine transform (DCT) is applied. The 13 first coefficients after the DC coefficient are extracted, and first and second order temporal derivatives are computed, creating a 39-dimension MFCC feature. Kraft et al. [25] modelize speech and non-speech features using Gaussian Mixture Models (GMM) with diagonal covariance matrix. In our experiments, we used 50 gaussians instead of 128 as proposed in [25], to avoid overfitting since we have less training data. When testing, the probability of each MFCC feature is computed for the speech and non-speech

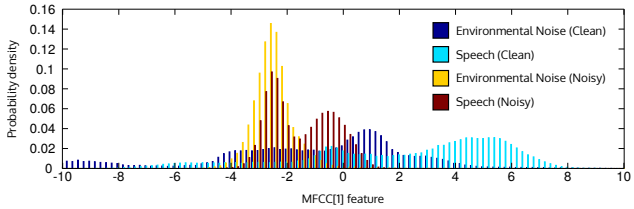


Fig. 3: Distribution of clean and noisy MFCC features (first coefficient) for speech and environmental noise

GMMs, and the sum of the log probabilities is computed over M frames. The block of frames is classified as speech if the sum computed using the speech GMM is greater than the one obtained with using non-speech GMM.

A true positive TP occurs when a speech signal is classified as speech, a true negative TN when a non-speech signals is classified as non-speech, a false negative FN occurs when a speech signal is classified as non-speech, and a false positive FP takes place when a non-speech signal is classified as speech. Accuracy is measured according to (11) [26]. An accuracy of 1.0 is desirable as it implies that the approach only generates true positives and true negatives.

$$\text{accuracy} = \frac{\sum TP + \sum TN}{\sum TP + \sum FP + \sum FN + \sum TN} \qquad (11)$$

Table II shows the performances of the proposed system and the MFCC-GMM system. When the SNR is high, performances of both system are similar and close to 100%, with the proposed approach being slightly more accurate in general. However, when the SNR is below 10 dB, the MFCC-GMM accuracy drops significantly, whereas the performance of the proposed system remain high. In some cases, such as when the reverberation time is 0 msec and the SNR is 0 dB with burst noise, the accuracy of the proposed method is more than twice the MFCC-GMM accuracy (100% vs 45%).

The accuracy of the proposed method is higher when SNR is low because the proposed pitch features are more robust than the MFCC features. Figure 3 illustrates the distribution of the first coefficient of the MFCC feature for the clean and noisy speech and environmental noise. Both speech and environmental noise MFCC distributions shift significantly when noise is added, and thus a pretrained model under clean conditions no longer models noisy distributions properly. Figure 4 shows the clean and noisy distribution of the pitch feature r for speech and environtmental noise. In this case, although the speech distribution slightly shifts towards zero, a clear separation between speech and environmental noise remains between $r = 0.1$ and $r = 0.2$.

B. Experiments on a Mobile Robot

The IRL-1 robot shown by Figure 5 is equipped with a eight-microphone array and the 8SoundUSB audio card [27] and was used for the trials. A male participant positioned himself at different locations around the robot and either spoke or produced a sound (hands clapping, coffee

TABLE II: Accuracy of the proposed system and MFCC-GMM (in parenthesis)

Noise type	SNR	Reverberation time (msec)				
		0	250	500	750	1000
White	20 dB	**100%** (99%)	**100%** (99%)	**100%** (99%)	**99%** (99%)	**99%** (99%)
	15 dB	**100%** (98%)	**100%** (98%)	**100%** (98%)	**99%** (97%)	**99%** (96%)
	10 dB	**100%** (97%)	**100%** (95%)	**100%** (95%)	**99%** (95%)	**99%** (95%)
	5 dB	**100%** (69%)	**99%** (60%)	**100%** (67%)	**99%** (71%)	**99%** (73%)
	0 dB	**100%** (50%)	**99%** (50%)	**99%** (50%)	**98%** (50%)	**98%** (51%)
Fan	20 dB	**100%** (99%)	**100%** (99%)	**100%** (99%)	99% (**100%**)	99% (**100%**)
	15 dB	**100%** (96%)	**99%** (98%)	**100%** (99%)	99% (**100%**)	99% (**100%**)
	10 dB	**100%** (98%)	**100%** (97%)	**100%** (98%)	**99%** (98%)	**99%** (98%)
	5 dB	**100%** (70%)	**100%** (61%)	**100%** (68%)	**99%** (72%)	**99%** (74%)
	0 dB	**100%** (50%)	**100%** (50%)	**100%** (50%)	**98%** (50%)	**98%** (51%)
Burst	20 dB	**100%** (99%)	**100%** (99%)	**100%** (99%)	**99%** (99%)	**99%** (99%)
	15 dB	**100%** (99%)	**100%** (99%)	**100%** (99%)	**99%** (99%)	**99%** (98%)
	10 dB	**100%** (99%)	**100%** (97%)	**100%** (97%)	**99%** (97%)	**99%** (98%)
	5 dB	**100%** (66%)	**100%** (58%)	**100%** (66%)	**99%** (70%)	**99%** (72%)
	0 dB	**100%** (45%)	**99%** (46%)	**100%** (48%)	**98%** (47%)	**98%** (48%)

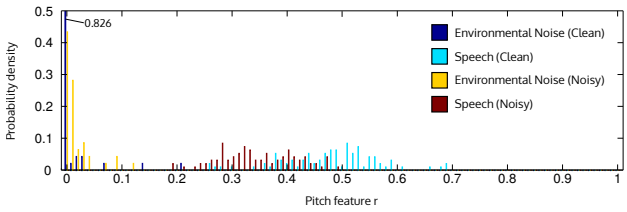


Fig. 4: Distribution of clean and noisy pitch features r for speech and environmental noise

cup clicked with a spoon, keys jangling, foot steps, phone ringing) for 10 sec at a time.

The reverberation level of the room was RT60 = 600 msec, and the SNR at each microphone varied between 10.7 dB and -5.5 dB (the noise mainly came from the fans onboard the robot, and their position relative to the microphones affected the SNRs differently). ManyEars' localization, tracking and separation modules found the azimut position of the source, and then the speech/non-speech discrimination module tagged the sound source as speech and non-speech. Figure 6a shows the position of each sound source and how it was classified (green for speech, red for non-speech). Figure 6b illustrates the tracked sources. A newly tracked source is initially shown in blue, and then changes color as it is classified as speech (green) or non-speech (red). Sound source tracking of the system could be improved to remove pauses (such as in situation A), but overall speech/non-speech discrimination is performed properly. The system performs classification after 2 sec of continuous tracking, and this explains the second classifications observed in segments B, C, D, E and F when tracking resumes after a pause. All classifications performed by the system in Fig. 6b match the type of sounds presented in Fig. 6a, which confirms the

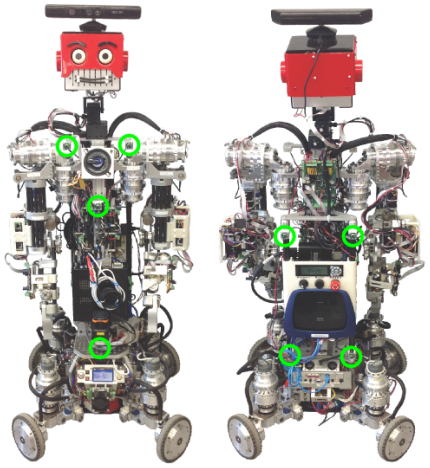
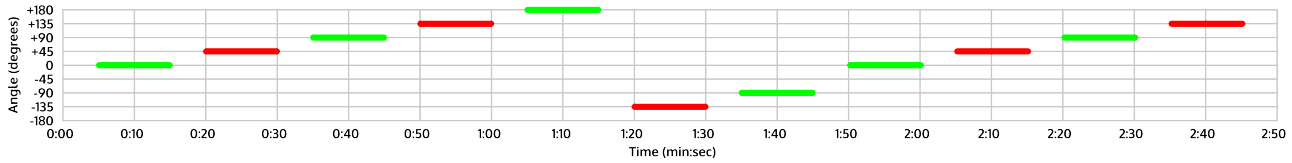


Fig. 5: IRL-1 robot with microphone locations identified by green circles

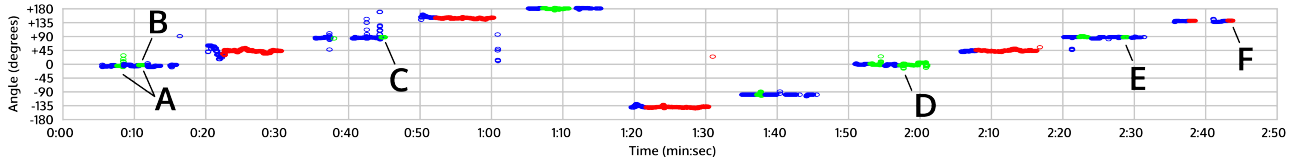
robustness and high accuracy of the proposed method.

VI. Conclusion

This paper presents a new speech/non-speech discrimination approach based on pitch estimation to make it robust to reverberation and additive noise. Simulations clearly demonstrate that the proposed system, which uses pitch features for speech/non-speech discrimination, is more robust to noise than a classification system based on the MFCC features and GMMs. Moreover, as opposed to the MFCC-GMM system, the approach works without *a priori* models of non-speech events, which makes it convenient for use on mobile robots operating in dynamic and changing environments. In future work, we plan to improve tracking and use the pitch estimation method to classify multiple sound sources simul-



(a) Theoretical positions of sound sources



(b) Tracked and classified sound sources

Fig. 6: Results of localization, tracking, separation and classification on IRL-1 (unclassified segments are shown in blue, speech in green and non-speech in red)

taneously active. We also want to study how pitch features can be used to improve speaker identification performance.

REFERENCES

[1] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, "The ManyEars open framework," *Autonomous Robots*, vol. 34, no. 3, pp. 217–232, 2013.

[2] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition HARK and its evaluation," in *Proc. of the IEEE Intl. Conf. on Humanoid Robots*, 2008, pp. 561–566.

[3] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array sources separation with post-filter," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, vol. 3, 2004, pp. 2123–2128.

[4] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Trans. Robotics*, vol. 23, no. 4, pp. 742–752, 2007.

[5] S. Yamamoto, K. Nakadai, J.-M. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and H. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proc. of the IEEE Intl. Conf. on Intelligent Robots and Systems*, 2005, pp. 4040–4045.

[6] F. Grondin and F. Michaud, "WISS, a speaker identification system for mobile robots," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2012, pp. 1817–1822.

[7] K. Murata, K. Nakadai, K. Yoshii, R. Takeda, T. Torii, H. G. Okuno, Y. Hasegawa, and H. Tsujino, "A robot singer with music recognition based on real-time beat tracking," in *Proc. of the Intl. Society for Music Information Retrieval*, 2008, pp. 199–204.

[8] A. Wang, "An industrial strength audio search algorithm," in *Proc. of the Intl. Society for Music Information Retrieval*, 2003, pp. 7–13.

[9] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound representation and classification benchmark for domestic robots," *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2014.

[10] Y. Sasaki, M. Kaneyoshi, S. Kagami, H. Mizoguchi, and T. Enomoto, "Daily sound recognition using pitch-cluster-maps for mobile robot audition," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2009, pp. 2724–2729.

[11] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Speaker identification under noisy environments by using harmonic structure extraction and reliable frame weighting," in *Proc. of INTERSPEECH*, 2006, pp. 1459–1462.

[12] J. Ramírez, P. Yélamos, J. M. Górriz, J. C. Segura, and L. Garcıa, "Speech/non-speech discrimination combining advanced feature extraction and SVM learning," in *Proc. of INTERSPEECH*, 2006, pp. 1662–1665.

[13] J. Ramírez, J.ez, J. M. Górriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," *IEEE Signal Processing Letters*, vol. 13, no. 8, pp. 497–500, 2006.

[14] R. Culebras, J. Ramírez, J. M. Górriz, and J. C. Segura, "Fuzzy logic speech/non-speech discrimination for noise robust speech Processing," in *Proc. of the Intl. Conf. on Computational Science*. Springer, 2006, pp. 395–402.

[15] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[16] R. Brueckmann, A. Scheidig, and H. Gross, "Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2007, pp. 1782–1787.

[17] M. Heck, C. Mohr, S. Stüker, M. Müller, K. Kilgour, J. Gehring, Q. B. Nguyen, V. H. Nguyen, and A. Waibel, "Segmentation of telephone speech based on speech and non-speech models," in *Speech and Computer*. Springer, 2013, pp. 286–293.

[18] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[19] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing,*, 2013, pp. 7204–7208.

[20] A. B. Nielsen, L. K. Hansen, and U. Kjems, "Pitch based sound classification," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2006, pp. 788–791.

[21] B. Bischl, U. Ligges, and C. Weihs, "Frequency estimation by DFT interpolation: A comparison of methods," Universität Dortmund, Tech. Rep., 2009.

[22] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[23] J. A. Bilmes *et al.*, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[25] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ICA for classification of acoustic events in a kitchen environment," in *INTERSPEECH, Lisbon, Portugal*, 2005.

[26] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, pp. 37–63, 2011.

[27] D. Abran-Côté, M. Bandou, A. Béland, G. Cayer, S. Choquette, F. Gosselin, F. Robitaille, D. Telly Kizito, F. Grondin, and D. Létourneau. (2012) EightSoundUSB. [Online]. Available: http://eightsoundsusb.sourceforge.net