Playful Interaction with Voice Sensing Modular Robots

Bjarke Heesche, Ewen MacDonald, Rune Fogh,
Moises Pacheco, and David Johan Christensen

Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

bjarke.heesche@gmail.com, {emcd, rufo, mpac, djchr}@elektro.dtu.dk

Abstract. This paper describes a voice sensor, suitable for modular robotic systems, which estimates the energy and fundamental frequency, F_0 , of the user’s voice. Through a number of example applications and tests with children, we observe how the voice sensor facilitates playful interaction between children and two different robot configurations. In future work, we will investigate if such a system can motivate children to improve voice control and explore how to extend the sensor to detect emotions in the user’s voice.

1 Introduction

We are developing the Fable modular robotic system to enable and motivate non-expert users to assemble and program their own robots from user-friendly building blocks [9]. A user assembles a robot by connecting several robotic modules together where each module provides functionality to perform a specific task, such as identifying users, characterizing the environment, moving, or manipulating its surroundings. In this paper we describe the development of a voice sensor for the Fable system that is able to measure different features (energy and F_0) of the user’s voice (in particular children’s). Our objectives with developing this sensor include the following:

1. Facilitate playful interaction between robots and users.
2. Enable users to create their own voice enabled robot applications.
3. Motivate users to improve the control of their voice (e.g., pitch and voice level).

In this paper, we focus on the first objective by demonstrating four applications that use the voice sensor to create games and playful interaction with the Fable system (see Fig. 1). In future work we will address the other two objectives by integrating the voice sensor and our programming tool-kit to enable the users to also program their own applications. As well, we plan to extend the sensor functionality to detect emotion in children’s voices.

Fig. 1. User testing with two Fable robots controlled using the voice sensor (voice energy and F_0 sensing). The humanoid has controllable arms and the quadruped has controllable locomotion.

2 Related Work

As the name implies, modular robotic systems consist of open-ended modules that can be combined in different ways to form a variety of different configurations [18, 19]. Our approach to designing modular playware [7] combines modular robotics with the concept of construction toys (such as LEGO) to facilitate playful creativity and learning. Examples of modular robotic playware include the Topobo system which enables the user to record and playback motor sequences using programming-by-demonstration [11] and the roBlocks (now Cubelets) which utilizes a programming-by-building strategy where the behavior emerges from the interaction of the modules with each other and their environment [15]. Similarly, LEGO Mindstorms, LEGO WeDo, and PicoCricket [13] are robotic construction kits that enable direct-user-programming of actuated and sensing models. In some ways, our Fable system is similar to these examples and is also influenced by Papert’s learning theory of constructionism [10]. Our aim is to enable everyone to become a robot designer and encourage them to imagine, create, play, share and reflect (i.e., the kindergarten approach to learning [12]).

We hypothesize that by enabling the creation of socially interactive robots, we can motivate a wider range of users to participate in this activity. In order to create these robots, we must develop sensors that can reliably interpret human communication. While, much research has been dedicated to machine recognition of human communication (e.g., [6]) and its applications to social robots (e.g., [2, 3, 14, 17]), our approach is to simplify the sensor design by limiting the detection to simple units of communication (e.g, gestures). Previous work along these lines to

track and detect a user’s posture using the Fable system has been promising [8]. In this paper, we limit the sensor to the detection of a single acoustic variable, the fundamental frequency or F_0 , and then examine if playful interaction can be sustained using input only from this detector.

3 The Voice Sensor

While the main focus of the sensor is on the estimation and tracking of the user’s F_0 over time, we must first detect if the user is vocalizing. Thus, the voice sensor consists of two elements: an energy detector and F_0 estimator.

3.1 Energy Detector

The simplest approach to detecting if a user is vocalizing is to measure and track in real-time the acoustic energy picked up by a microphone. A simple approach is to calculate the root mean square (RMS) of the signal over some window. However, to reduce the number of computations, the signal is squared and low-pass filtered using an exponential moving average (EMA). Thus, the energy E , for a given signal $x[i]$ is estimated as follows:

$$y[i] = \alpha \times y[i - 1] + (1 - \alpha) \times x[i]^2 \quad (1)$$

where α is a weighting factor. For our application, with a sampling frequency, F_s of 8 kHz, we set $\alpha = 0.990$. Thus, the time constant of the EMA was 15 ms and resulted in a relatively stable measure of E during vocalizations. The output is energy in dB relative to $E_{ref} = 1$ is as follows:

$$E_{dB}[i] = 10 \times \log_{10}(y[i]) \quad (2)$$

If a signal is detected with an energy level, $E[i] > \beta$, where $\beta = 0.05$, it is assumed to be a sound produced by the user.

3.2 F_0 Sensor

In speech, F_0 refers to the fundamental frequency, which corresponds to the frequency at which the vocal folds open and close. However, not all speech is voiced (i.e., uses the vocal folds to produce a periodic acoustic excitation). If the vocal tract is sufficiently constricted, then the air flow will become turbulent, which produces sound. As well, for plosive sounds, airflow can be stopped and the pressure can be allowed to build significantly before the air is released in a burst. Thus, speech can be separated into voiced (periodic) or un-voiced (aperiodic) sounds. However, when estimating F_0 , we currently assume that the signal is always voiced. In the future, we plan to add a harmonicity detector to determine if the vocalizations produced are voiced or unvoiced.

As with the energy detector, the sensor is designed to estimate F_0 in near-real-time (i.e., low latency) to facilitate interaction. Further, as the goal is to

Fig. 2. Elements in the F_0 sensor.

Fig. 3. Elements in F_0 estimation.

develop inexpensive modular sensors, the algorithm used should be feasible to implement using current embedded processors for autonomous control.

The F_0 sensor consists of three major elements (see Fig. 2). The energy detector, described previously, is used to determine if the user is vocalizing. If speech is detected, the second element estimates F_0 of the microphone signal. The final element, F_0 correction, monitors the F_0 estimates to detect and removes outliers caused by noise or unvoiced sounds.

F_0 Estimation: The F_0 estimator uses an autocorrelation method based on [4], which is simple and computationally efficient (see Fig. 3). The autocorrelation function $r_t(\tau)$ is computed for short time windows of length W_{auto} . If the signal is periodic then the autocorrelation function will have peaks at lags τ corresponding to the period of the signal. Thus, the fundamental frequency is estimated by finding the time lag corresponding to the first peak in the autocorrelation function. To improve performance in picking the correct peak (i.e., the one corresponding to the fundamental frequency), two modifications are made.

First, the autocorrelation function is multiplied by a decay function.

$$r_t(\tau) = \sum_{(j=t+1)}^{(t+W_{auto})} [x_j x_{j+\tau} (1 - \tau/\tau_{max})] \quad (3)$$

Second, the search for the peak is restricted to the interval $\tau \in [\tau_{min} : \tau_{max}]$. Thus, by restricting the range of possible lags, the range of F_0 estimates is also restricted. While the average fundamental frequency for a child is 300 Hz for speech [5], F_0 can be higher for non-speech vocalizations such as humming. Thus, to cover the accessible frequency spectrum while keeping the computational costs low, we set τ_{min} to correspond with an F_{max} of 800 Hz. To allow the sensor to also track adults, τ_{max} was set to correspond with an F_{min} of 44 Hz. To further reduce computational costs, the modified correlation function was only computed over the interval $\tau \in [\tau_{min} : \tau_{max}]$.

As the sampling frequency of the signal is increased, the computational costs of the autocorrelation function increase significantly. Thus, the sampling rate of the sensor, F_s , was set at 8 kHz. However, this low sampling rate resulted in poor precision in F_0 estimation, particularly at higher frequencies. Thus, the modified autocorrelation function was interpolated before searching for the maximum peak.

In summary, the output from the F_0 estimator is the frequency that corresponds with the lag of the maximum peak of the modified autocorrelation function over the interval $\tau \in [\tau_{min} : \tau_{max}]$. Using sinusoidal test signals, an effective estimation range of 120–800 Hz was found. This covers the range of frequencies we would expect for children’s, adult women’s, and most adult men’s vocalizations.

F_0 Correction: While the F_0 estimator works well, some artefacts are observed in running speech due to unvoiced speech sounds or when the sensor is placed in a noisy environment. Thus, an F_0 correction stage was developed.

For outlier detection, the F_0 estimates are monitored and the interquartile range (*IQR*; i.e., range between the first and third quartile) is estimated. Using this, upper and lower boundaries are set to $Q_3 + 1.5 \times IQR$ and $Q_1 - 1.5 \times IQR$ respectively. Estimates outside this range are considered to be outliers. Before discarding the outlier, an octave correction is tested. The outlier is shifted up or down by an octave towards the frequency range where most F_0 estimates are present. If the difference between two neighbour estimates is less than 10 Hz, the correction is considered to be successful. Otherwise, the outlier is discarded.

Performance: An example of F_0 estimation is shown in Fig. 4. To evaluate the performance of the system, 1150 children’s utterances [16] were tested. Each utterance was processed by both the voice sensor and Praat [1] and the F_0 estimates were compared. As there were occasions where Praat detected voiced speech but the sensor did not (and vice versa), the comparison was restricted

Fig. 4. Example of voice and F0 estimation for a speech signal.

to time points where both Praat and the voice sensor produced F_0 estimates. Overall, when F_0 correction was not used in the voice sensor, the mean error was 14 Hz. When F_0 correction was employed, the mean error was reduced to 4 Hz. In summary, the system is sufficiently accurate for our applications, but we can improve its performance in future work.

4 Test of Applications

4.1 Test Setup

A total of four different applications were tested, across two configurations of the Fable system (one using a humanoid and three using a quadruped; see Fig. 1). Children’s vocalizations were recorded using headsets connected to a computer. The voice sensor processing was conducted in Matlab running on a standard PC, which sent commands over a wireless serial connection to control the Fable actuators.

Four girls and four boys ranging in age from 7-10 years participated in the testing. The children were divide into four pairs according to their age: younger girls (7-8 years old), older girls (10 years old), younger boys (8-9 years old) and older boys (10 years old). As the average F_0 can vary across age and gender the individual parameters $F_{0,min}$ and $F_{0,max}$ were measured for each child prior to the test.

The aim of the tests was to demonstrate if the Fable system and the voice sensor can be successfully used by children to create playful interactions. The tests were designed as small games in order to motivate the users to engage and play with the system.

Fig. 5. F_0 control of humanoid Fable. The user controlled arm changes its position according to the F_0 produced by the user. Once the target F_0 is produced, the robot starts waving both hands.

4.2 Humanoid Fable Application

The first application was developed to provide an introduction to how the users could control the robot with their voice. In the game, the goal was to angle both of the Fable arms the same way (see Fig. 5). One arm, the target, was initially angled by the experimenter using a gamepad. The other arm was controlled by the F_0 of the user's voice. The user hummed or sung a tone into the microphone and the arm would raise or lower based on the F_0 that was produced. Once the controlled arm matched the target, the Fable started waving its arms. This waving continued for as long as the user could hold the tone. This task was selected as a good introduction because the children received clear visual feedback from the robot (i.e., the angle of its arm) that corresponded to the pitch they produced. Three different random target values were tried for each of the eight children.

The task was easily understood and all of the children were able to complete it. However the older children were slightly better at the task. Initially, two of the younger children had difficulties because they were changing their pitch too fast and/or in large steps. As well, the younger girls required more instruction before understanding the relationship between their voice and Fable's arm position. Regardless, the most common strategy for completing the task was to start at a random pitch and adjust it up or down seeking the target position. Even though the task was developed as an introduction to the concept of controlling a robot by voice, the children were very engaged in the activity and appeared to be enjoying themselves.

4.3 Quadruped Fable Applications

In the following applications the objective was to move the quadruped robot through a track with four gates. The F_0 of the users' voices was used to select between four different movements (forward, backward, left-turn, right-turn). In addition to the robot's movement, visual feedback was also provided on a computer screen. The F_0 being produced was plotted as a large dot that moved over a background that illustrated the target pitch corresponding to each motion. Three different applications were tested:

Fig. 6. F_0 control of quadruped Fable. By producing an F_0 that corresponds to one of the targets, the user can select different movements and control the robot.

Single Player: In this game, there were two different levels of difficulties. At level 1, there were only two pitch targets corresponding to forward and left-turn movements. At level 2, there were four pitch targets, each corresponding to one of the four movements (see Fig. 6).

Two Player - Collaborative: In this game, two players collaborated to control the quadruped using two pitch targets. One user controlled forward and backward movement while the other controlled turning left and right.

Two Player - Competitive: In this game, two players competed to gain control over the quadruped and move it to a specific area to win the game. One user controlled forward and left-turn and the other user controlled backward and right-turn. The user that was loudest, while maintaining a target pitch, controlled the robot. However, to avoid the users screaming as loud as possible, a maximum value was set for the voice energy. If the user exceeded this level, control was lost until the voice level was reduced below the maximum threshold.

All eight children participated in the single player game at level 1. The four oldest quickly understood the task and learned how to shift between tones in order to get the robot moving through the track. In contrast, only one of the younger boys performed well. The others needed many attempts before completing the task. The two older boys tried level 2 and were able to control the robot successfully. While we anticipated that level 2 would be more difficult, the boys were able to complete the route faster than at level 1.

As the younger children had difficulty with the singleplayer game it was decided that they would participate in the collaborative two-player game with the older children. The collaborative game added an additional complexity as the users needed to communicate with each other without triggering the robot by mistake. This challenge engaged the children to communicate with gestures or mute the microphone, and they were surprisingly good at completing the task. One of the boys commented that the single player game on level 1 was a bit hard, because he could only turn left. He also felt that level 2 was both easier

and harder. It was easier because he had more control options, but it was more difficult because of all the target pitches he had to remember. However, he felt the collaborative game was just right, and really fun, because they had access to all the four actions, but only had to remember two notes at a time.

The competitive two-player game was only tested with the older boys. This application turned out to be the most difficult of the four tested. It was difficult for the boys to control the tone and the power of their voice simultaneously. As the position of the microphone significantly influences the energy measurements of the voice, one of the boys adjusted the microphone position rather than the level of his voice.

Across all the applications, the younger children had difficulties holding a specific pitch for a long period of time. Instead, they would hum a small melody. This approach might have been easier for them had they been instructed to change between two melodies instead of two tones. As a result, the older children seemed to have more fun when playing with the system. The two older boys continuously asked to play the collaborative game again after finishing. If the interactions are not changed, it is suggested that future tests focus on the older age group (i.e., 9 years and older). In summary, these example applications confirmed that we can use the voice sensor to facilitate playful interaction between the children and Fable robots. In addition, we believe that the children were also learning to better control their voices but further work is needed to test this.

5 Conclusion and Future Work

This paper presented a voice sensor for the Fable modular robotic system. The sensor measures the energy in the user’s voice and estimates F_0 , which is related to the pitch of the voice, with very low latency. The performance of the F_0 estimation was tested and found sufficient for the purpose. Four different applications utilizing the voice sensor were tested with eight children and we found that the system facilitated interaction that was challenging, playful and motivating.

In future work, more voice sensor functionality will be added and the possibility of extending the current processing for use in emotion detection will be explored. In addition, we will integrate the voice sensor with our existing software tool-kit to enable users to program their own applications.

6 Acknowledgements

This work was performed as part of the “Modular Playware Technology” project funded by the Danish National Advanced Technology Foundation.

References

1. P. Boersma and D. Weenink. Praat [computer software]. *Amsterdam: University of Amsterdam*, 2006. <http://www.fon.hum.uva.nl/praat/>.

2. C. Breazeal and L. Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous robots*, 12(1):83–104, 2002.
3. C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. *rm*, 255:3, 1999.
4. A. De Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002.
5. F. Jacobsen, T. Poulsen, J. H. Rindel, A. Christian Gade, and M. Ohlrich. Fundamentals of acoustics and noise control, 2010. Note no 31200.
6. R. P. Lippmann. Speech recognition by machines and humans. *Speech communication*, 22(1):1–15, 1997.
7. H. H. Lund. Modular interactive tiles for rehabilitation: evidence and effect. In *Proceedings of the 10th WSEAS international conference on Applied computer science*, ACS’10, pages 520–525, Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS).
8. A. Magnusson, M. Pacheco, M. Moghadam, H. H. Lund, and D. J. Christensen. Fable: Socially interactive modular robot. In *Proceedings of 18th International Symposium on Artificial Life and Robotics*, Daejeon, Korea, January 2013.
9. M. Pacheco, M. Moghadam, A. Magnusson, B. Silverman, H. H. Lund, and D. J. Christensen. Fable: Design of a modular robotic playware platform. In *Proceedings of the IEEE Int. Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, May 2013.
10. S. Papert. *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc., 1980.
11. H. S. Raffle, A. Parkes, and H. Ishii. Topobo: a constructive assembly system with kinetic memory. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI ’04, pages 647–654, New York, NY, USA, 2004. ACM.
12. M. Resnick. All i really need to know (about creative thinking) i learned (by studying how children learn) in kindergarten. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition*, pages 1–6. ACM, 2007.
13. M. Resnick and B. Silverman. Some reflections on designing construction kits for kids. In *Proceedings of the 2005 conference on Interaction design and children*, pages 117–122. ACM, 2005.
14. J. Saldien, K. Goris, S. Yilmazyildiz, W. Verhelst, and D. Lefeber. On the design of the huggable robot probro. *Journal of Physical Agents*, 2(2):3–12, 2008.
15. E. Schweikardt and M. D. Gross. roblocks: a robotic construction kit for mathematics and science education. In *Proceedings of the 8th international conference on Multimodal interfaces*, ICMI ’06, pages 72–75, New York, NY, USA, 2006. ACM.
16. S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children’s Speech*. PhD thesis, University of Erlangen-Nuremberg, 2009.
17. W. D. Stiehl, J. K. Lee, C. Breazeal, M. Nalin, A. Morandi, and A. Sanna. The huggable: a platform for research in robotic companions for pediatric care. In *Proceedings of the 8th International Conference on Interaction Design and Children*, IDC ’09, pages 317–320, New York, NY, USA, 2009. ACM.
18. K. Stoy, D. Brandt, and D. J. Christensen. *Self-Reconfigurable Robots: An Introduction*. Intelligent Robotics and Autonomous Agents series. The MIT Press, 2010.
19. M. Yim, B. Shirmohammadi, J. Sastra, M. Park, M. Dugan, and C. J. Taylor. Towards robotic self-reassembly after explosion. In *Video Proceedings of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, San Diego CA, 2007.