A Dynamic Speech Breathing System for Virtual Characters

Ulysses Bernardet[1] ✉, Sin-hwa Kang[2], Andrew Feng[2], Steve DiPaola[1], and Ari Shapiro[2]

[1] School of Interactive Arts and Technology, Simon Fraser University, Vancouver, Canada
{ubernard,sdipaola}@sfu.ca
[2] University of Southern California, Institute for Creative Technologies, Los Angeles, California
{kang,feng,shapiro}@ict.usc.edu

Abstract. Human speech production requires the dynamic regulation of air through the vocal system. While virtual character systems commonly are capable of speech output, they rarely take breathing during speaking – speech breathing – into account. We believe that integrating dynamic speech breathing systems in virtual characters can significantly contribute to augmenting their realism. Here, we present a novel control architecture aimed at generating speech breathing in virtual characters. This architecture is informed by behavioral, linguistic and anatomical knowledge of human speech breathing. Based on textual input and controlled by a set of low- and high-level parameters, the system produces dynamic signals in real-time that control the virtual character's anatomy (thorax, abdomen, head, nostrils, and mouth) and sound production (speech and breathing). The system is implemented in Python, offers a graphical user interface for easy parameter control, and simultaneously controls the visual and auditory aspects of speech breathing through the integration of the character animation system SmartBody [16] and the audio synthesis platform SuperCollider [12]. Beyond contributing to realism, the presented system allows for a flexible generation of a wide range of speech breathing behaviors that can convey information about the speaker such as mood, age, and health.

Keywords: Speech breathing, speaking, breathing, virtual character, animation

1 Introduction

In animals, breathing is vital for blood oxygenation and centrally involved in vocalization. What about virtual characters? Does the perceivable presence or absence of this behavior that is so vital in biological systems play a role in how they are perceived? Is breathing movement, frequency, sound etc. effective at conveying state and trait related information? These are some of the questions that motivate the research into breathing in virtual characters presented here.

Breathing is a complex behavior that can be studied both, on its own, and in relation to other behaviors and factors such as emotion and health. In the work we present here, we focus our interest on the dynamic interplay between speaking and breathing, on what is called "speech breathing". From a functional perspective, the respiratory system needs to provide the correct pressure drive to the voice box [10]. Consequently, breathing is implicated in many aspects of speech production [20] such as voice quality, voice onset time, and loudness.

1.1 Related work

Breathing in virtual human characters As [18] point out, the more realistic virtual characters are becoming overall, the more important it is that the models are realistic at the detail level. Though the importance of including the animation of physiological necessities such as breathing has been recognized [14], few virtual character systems actually take breathing into account. In their interactive poker game system, [6] include tidal breathing – inhalation and exhalation during restful breathing – of the virtual characters as a means of expressing the character's mood. Models of tidal breathing that strive to be anatomically accurate include those developed by [18,21]. Recent models are usually based on motion data captured from participants [15,17]. Modeling work on breathing in conjunction with speaking is sparse, and the work of [9] on the development of infant speech production one of the few published works.

Physiology of (speech) breathing A number of experimental studies have investigated the relationship between the two processes of breathing and speaking. Empirical research has shown that the respiratory apparatus is sensitive to the demands of an upcoming vocal task, and that kinematic adjustments take place depending on where speech was initiated within the respiratory cycle [13]. The two distinct parts of the speech breathing process are the filling of the lungs referred to as "inhalation" or "inspiration", and the outflow of air – exhalation or expiration – that drives the voice box. Figure 1 shows the key parameters relating to the dynamics and interplay between breathing and speaking. Inspiration normally takes places at construct boundaries such as at the end of a sentence [7,8,19]. The key parameter pertaining to expiration is the utterance length, i.e. the number of syllables or words produced on one speech breath. In speech breathing one cycle of inspiration and expiration is referred to as "breath group" (Figure 1). In their study [19] found that 'breath group' lengths during both, reading and spontaneous speech tasks, had a duration of 3.84 seconds with a rather large standard deviation of 2.05 seconds. While relatively stable within a single participant, larger differences, ranging from 0.3 to 12.6 seconds, were found between participants.

In this paper, we present our work on a dynamic speech breathing model. The system consists of several tightly synchronized processes for inspiration and expiration animation, inspiration sound, audible speech, and facial animations (lip synch, mouth open, and nostril fare). All behavior is generated in real-time

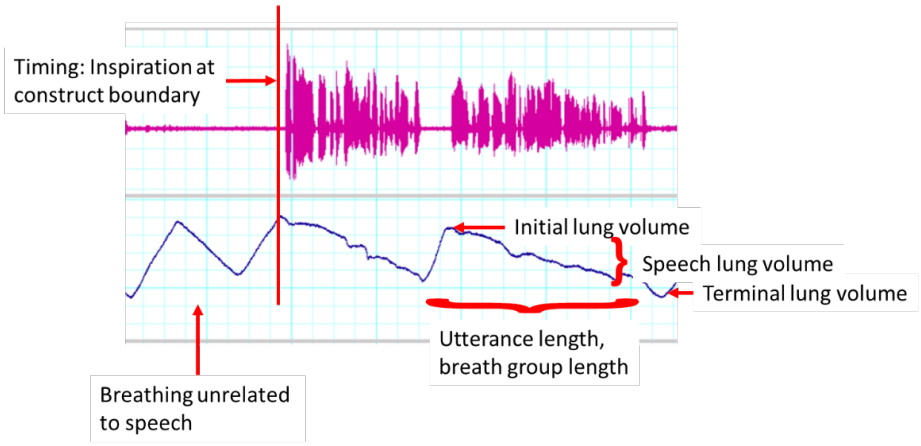


Fig. 1: Dynamics of speech breathing

and controllable via a set of low- and high-level parameters. Our goal is to provide a dynamic and tunable speech breathing system that contributes to augmenting the realism of computer generated virtual humanoid characters.

2 Dynamic speech breathing system

2.1 System overview

The open input to the system is the text, while tunable inputs are parameters controlling the speech and the breathing processes. At the output side, generated by the *control model*, the system produces dynamic control signals for shapes (thorax, abdomen, head, nostrils, and mouth) and sounds (speech and breathing).

2.2 Control model

At the core of the speech breathing control model stands the oscillation between two fundamental processes: inspiration and expiration (Figure 2).

Inspiration process Physiologically, the function of inspiration is filling the lungs with air. In our model, inspiration comprises four independent, parallel processes: Triggering of facial animations (mouth and nostrils), inspiration animation (thorax, abdomen, neck), playback of inspiration sounds (see implementation section for details), and speech preparation. $length_{inspiration}$ is only the tunable parameter for the inspiration process. It is an independent parameter, because, based on what is know about physiological processes, the length of the inspiration is mostly independent of both, the length of the utterance and the lung volume. The inspiration animation consists of breathing-induced shape changes to

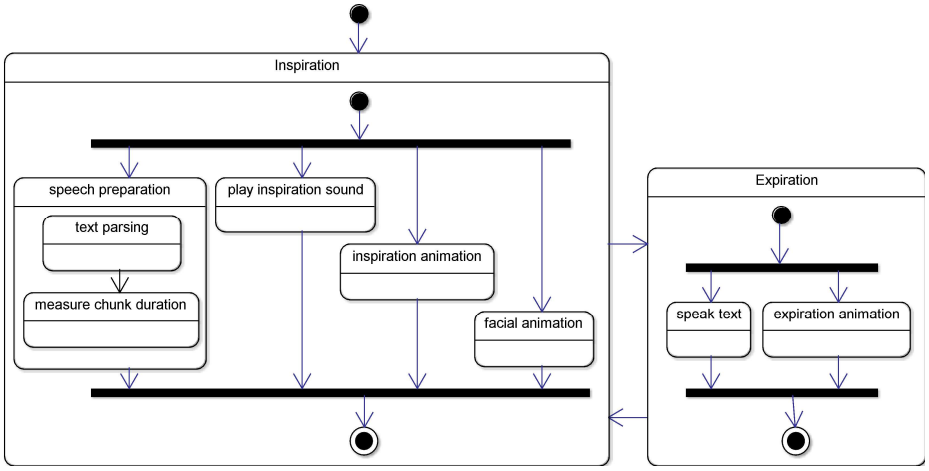


Fig. 2: State diagram of the dynamic breathing model

the chest and stomach, as well as a pitch rotation of the head along the coronal axis. All three of these changes are driven by a linear function LF with a slope defined as $\frac{volume_{lung}}{length_{inspiration}}$. The variable $volume_{lung}$ in turn, is a function of the length of the upcoming speech output (for details see "speech preparation process" below). Maximal volumetric and angular changes are set through the parameters $volume_{thorax}$, $volume_{abdomen}$, and $amplitude_{neck\ motion}$, respectively. Additionally, the model controls the change to two other shapes: The flaring of the nostrils, and the opening of the mouth. The maximal amplitudes for these are set by, $amplitude_{nostril\ flare}$, and $amplitude_{mouth\ open}$, respectively.

The system can produce two different inspiration sounds; breathing through the mouth and breathing through the nose. The Loudness of these two sound types is controlled by the parameters $amplitude_{sound\ mouth}$, and $amplitude_{sound\ nose}$, respectively. For clarity, we use the term "loudness" when referring to sound amplitude, and "volume" when referring to volumetric entities such as lungs.

Parallel to these processes, which produce perceivable output, runs the speech preparation process. Speech preparation comprises two steps. In a first step, the input text is parsed to extract the text chunk that will be spoken. The following steps define the text parsing algorithm (also see Figure 3):

- Step through text until number of syllables specified by the $length_{utterance}$ parameter is reached
- Map the position back onto the original text
- Search text forward and backward for the position of "pause markers" period (".") and underscore ("_")
- If the position of both pause markers (in number of characters) is larger than the parameter $urgency_{limit}$, define pause at word boundary
- Otherwise, define pause at position of pause marker, with priorities "." > "_"

– Identify text chunk for utterance and set remaining text as new input text

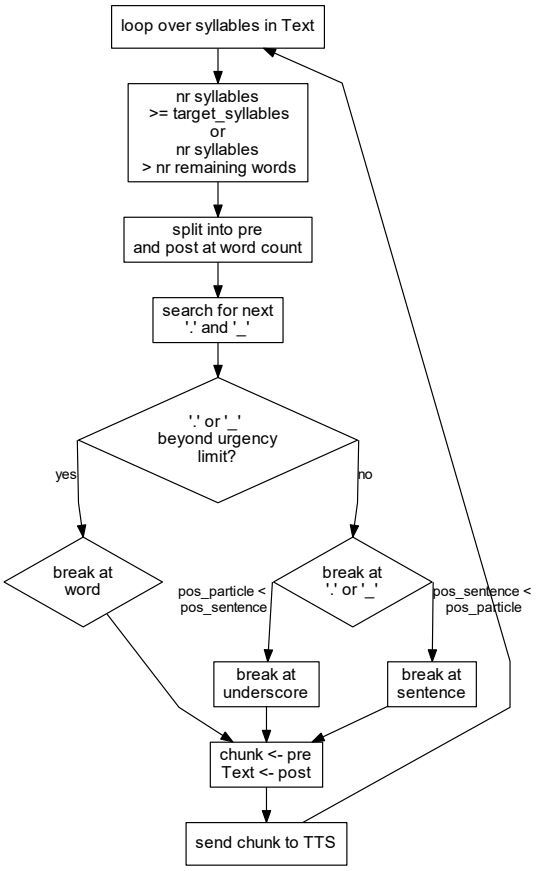


Fig. 3: Flow chart of the text parsing process

Note that we introduce the concept of "pause markers" to be able to have a more fine-grain control of the speech breathing process. The $urgency_{limit}$ parameter effectively defines how much flexibility the model has in terms of deciding when to insert inspiration into the text stream (see detailed explanation below).

The second step of the speech preparation process is the measurement of the upcoming speech output length (in seconds). This is done in an empirical fashion by sending the text to the text-to-speech system (TTS) and measuring the length of the generated audio bitstream. This approach is necessary because the actual length of the spoken text depends on the speech parameters, e.g. rate, as well as on the specifics of the text-to-speech system, e.g. the voice used.

Expiration process Two parallel processes make up the expiration phase: The generation of the stream of spoken output by the TTS and the expiration animation. The two parameters directly controlling the speech production are $prosody_{rate}$ and $prosody_{loudness}$. As for inspiration, a linear function controls thorax, abdomen, and neck. However, in the expiration case, the slope of the function is defined as $\frac{volume_{lung}}{length_{speech}}$.

The output of the oscillation between the inspiration and expiration process, as well as the speech and breathing sounds is illustrated in Figure 4.

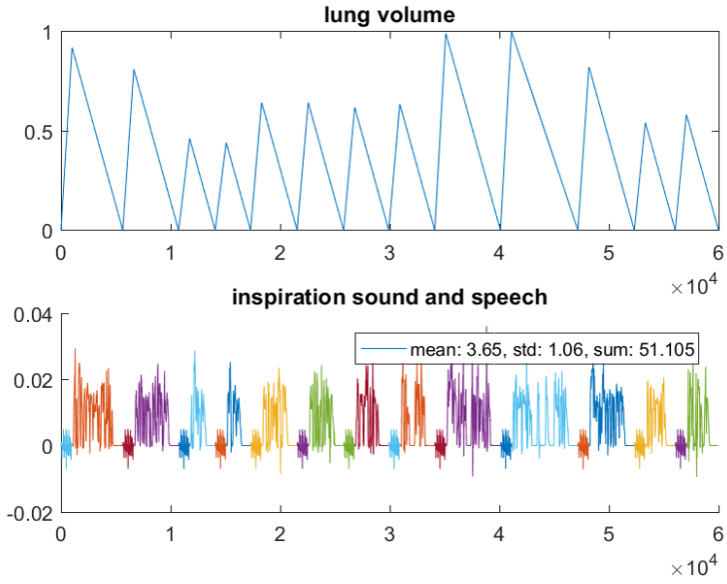


Fig. 4: Time course of the model behavior and outputs. The top saw-tooth function is used to control the thorax and abdomen shapes, as well as the pitch angle of the head. The lower panel shows the speech and breathing sound output of the model.

Abstract control Tuning individual low-level parameters is not desirable in most applications; rather, we would like to have a more abstract control model with a limited set of parameters. Additionally, the abstract control model ensures that low-level parameters are in a sensible causal relationship. While we subsequently lay out the parameters of the model, Equation 1 shows the qualitative model. Two parameters are related "breathing style": *ThoraxVsAbdomen* defines how much the character is chest or stomach breathing, while *MouthVsNose* controls inspiration through mouth vs. nose. The overall capacity of the lung is defined by $capacity_{lung}$; The parameter $amplitude_{breathing\ sound}$ controls the overall

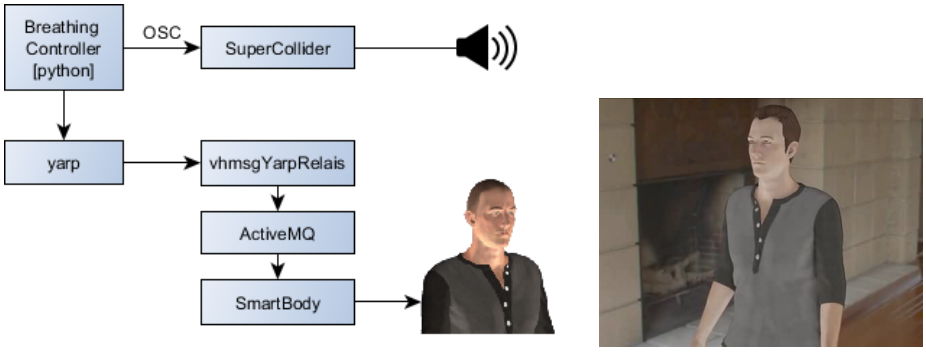
loudness of the inspiration sound, while $opening_{inspiration\ channels}$ the "inspiration channels" are opened. Low-level parameters that remain independent are $speaking_{loudness}$, $prosody_{rate}$, $length_{inspiration}$, and $amplitude_{neck\ motion}$.

$$
\begin{aligned}
\text{amplitude}_{\text{sound nose}} &= \text{MouthVsNose} * \text{amplitude}_{\text{breathing sound}} \\
\text{amplitude}_{\text{sound mouth}} &= (1 - \text{MouthVsNose}) * \text{amplitude}_{\text{breathing sound}} \\
\text{amplitude}_{\text{nostril flare}} &= \text{MouthVsNose} * \text{opening}_{\text{inspiration channels}} \\
\text{amplitude}_{\text{mouth open}} &= (1 - \text{MouthVsNose}) * \text{opening}_{\text{inspiration channels}} \\
\text{volume}_{\text{abdomen}} &= \frac{\text{ThoraxVsAbdomen} * \text{capacity}_{\text{lung}}}{100} \\
\text{volume}_{\text{thorax}} &= \frac{(1 - \text{ThoraxVsAbdomen}) * \text{capacity}_{\text{lung}}}{100} \\
\text{length}_{\text{utterance}} &= \sqrt{\frac{\text{capacity}_{\text{lung}}}{\text{speaking}_{\text{loudness}}}} * \text{norm}_{\text{syllables}} \\
\text{length}_{\text{inspiration}} &= \frac{\text{capacity}_{\text{lung}}}{100} \\
\text{urgency}_{\text{limit}} &= \text{length}_{\text{utterance}} * 2
\end{aligned}
\tag{1}
$$

2.3 Implementation

Breathing sounds The nose and mouth breathing sounds were recorded from one of the authors using a Rode NT1-A microphone, iRig PRE preamplifier, and Audacity software [1]. Post recording, the sound files were normalized and lengthen to five seconds by applying a Paulstretch filter using Audacity. During run-time, the sounds are played back using the audio synthesis and algorithmic composition platform SuperCollider [12]. The amplitude and length of the play back are controlled by applying a trapezoid envelope function to each of the waveforms (nose and mouth sound).

Real-time control architecture The core controller of the system is implemented in the Python programming language. The control commands for the sound playback are sent to SuperCollider via the Open Sound Control (OSC, [2]). Concurrently, the controller, via the 'm+m" middleware software [5], sends messages to the SmartBody virtual character system, where the character animation and rendering take place [16]. Thorax, abdomen, as well as facial animations, are implemented using blendshapes, while the head and mouth are controlled at the level of joint-angles. From within SmartBody, control signals are sent to the text-to-speech system (Microsoft TTS with the "Adam" voice from CereProc [3]. To facilitate parameter tuning and control, we developed a Graphical User Interface that was designed using Pygubu [4] and is based on Python's Tkinter module.



(a) (b)

Fig. 5: (a) Software architecture of the speech breathing control system (b) Screen capture of the virtual character controlled and rendered using the SmartBody system [16]

3 Discussion and Conclusion

We present a real-time control system for speech breathing in virtual characters, that is based on empirical knowledge about human speech behavior and physiology. The system receives input text and produces dynamic signals that control the virtual character's anatomy (thorax, abdomen, head, nostrils, and mouth) and sound production (speech and breathing). At the core of the speech breathing control model stands the oscillation between inspiration and expiration. The independent control of the physiologically grounded speech parameters allows the system to produce in real-time a wide range of speech breathing behaviors. In its present form, the control system offers the ability to produce physiologically grounded parameters of speech, e.g. utterance length function of speaking loudness and lung capacity.

The biggest limitation of the control system at this moment is the delicacy of the timing. A system intrinsic fragility stems from measuring the utterance length during the inspiration phase; if this measurement takes longer than the inspiration, the subsequent temporal coordination is compromised. An extrinsic source of potential desynchronization is the lag of speech onset in the Text-To-Speech system. A second limitation is that the TTS does not allow for variations in pitch. Especially pitch declination over an utterance, which might be related to subglottal pressure, and hence breathing [11] might be important for realism.

Next steps in the development of the system include the empirical investigation of the effect speech breathing has on factors such as realism and relatability. At the technical level, future steps include the improvement of the breathing animation to a state of the art implementation as presented e.g. in [17]. At the conceptual level, the system will be extended to include the control of speech breathing parameters with the goal of generating the expression of abstract constructs such as personality and emotion. For example, fast pace and shallow

breathing may lead to the perception of anxiety, while long and deep breathing may lead to the perception of calmness.

Acknowledgements

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant (No. R0184-15-1030, MR Avatar World Service and Platform Development using Structured Light Sensor) funded by the Korea government (MSIP).

References

1. http://www.audacity.audio/
2. http://opensoundcontrol.org/
3. http://www.cereproc.com/
4. https://github.com/alejandroautalan/pygubu
5. Bernardet, U., Schiphorst, T., Adhia, D., Jaffe, N., Wang, J., Nixon, M., Alemi, O., Phillips, J., DiPaola, S., Pasquier, P.: m+m: A Novel Middleware for Distributed, Movement Based Interactive Multimedia Systems. In: Proceedings of the 3rd International Symposium on Movement and Computing - MOCO '16. pp. 1–21. ACM Press, New York, New York, USA (2016), http://dl.acm.org/citation.cfm?doid=2948910.2948942
6. Gebhard, P., Schröder, M., Charfuelan, M., Endres, C., Kipp, M., Pammi, S., Rumpler, M., Türk, O.: IDEAS4Games: Building expressive virtual characters for computer games. In: Lecture Notes in Computer Science. vol. 5208 LNAI, pp. 426–440 (2008)
7. Henderson, A., Goldman-Eisler, F., Skarbek, A.: Temporal Patterns of Cognitive Activity and Breath Control in Speech. Language and Speech 8(4), 236–242 (1965)
8. Hixon, T.J., Goldman, M.D., Mead, J.: Kinematics of the Chest Wall during Speech Production: Volume Displacements of the Rib Cage, Abdomen, and Lung. Journal of Speech Language and Hearing Research 16(1), 78 (3 1973), http://jslhr.pubs.asha.org/article.aspx?doi=10.1044/jshr.1601.78
9. Howard, I.S., Messum, P.: Modeling Motor Pattern Generation in the Development of Infant Speech Production. 8th International Seminar on Speech Production pp. 165–168 (2008)
10. Huber, J.E., Stathopoulos, E.T.: Speech Breathing Across the Life Span and in Disease. The Handbook of Speech Production pp. 11–33 (2015), http://dx.doi.org/10.1002/9781118584156.ch2
11. Ladd, D.R.: Declination: a review and some hypotheses. Phonology 1(1), 53–74 (1984)
12. McCartney, J.: Rethinking the Computer Music Language: SuperCollider. Computer Music Journal 26(4), 61–68 (12 2002), http://www.mitpressjournals.org/doi/10.1162/014892602320991383
13. McFarland, D.H., Smith, A.: Effects of vocal task and respiratory phase on prephonatory chest wall movements. Journal of speech and hearing research 35(5), 971–82 (10 1992), http://www.ncbi.nlm.nih.gov/pubmed/1447931
14. Rickel, J., Information, U.S.C., André, E., Badler, N., Cassell, J.: Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intelligent Systems (2002)

15. Sanders, B., Dilorenzo, P., Zordan, V., Bakal, D.: Toward Anatomical Simulation for Breath Training in Mind/Body Medicine. In: Magnenat-Thalmann, N., Zhang, J.J., Feng, D.D. (eds.) Recent Advances in the 3D Physiological Human. Springer (2009), http://graphics.cs.ucr.edu/papers/sanders:2008:TAS.pdf
16. Shapiro, A.: Building a character animation system. In: Motion in Games. pp. 98–109 (2011)
17. Tsoli, A., Mahmood, N., Black, M.J.: Breathing life into shape. ACM Transactions on Graphics 33(4), 1–11 (7 2014)
18. Veltkamp, R.C., Piest, B.: A Physiological Torso Model for Realistic Breathing Simulation. In: Proceeding 3DPH'09 Proceedings of the 2009 international conference on Modelling the Physiological Human. pp. 84–94 (2009)
19. Winkworth, A.L., Davis, P.J., Adams, R.D., Ellis, E.: Breathing patterns during spontaneous speech. Journal of speech and hearing research 38(1), 124–144 (2 1995), http://www.ncbi.nlm.nih.gov/pubmed/7731204
20. Włodarczak, M., Heldner, M., Edlund, J.: Breathing in Conversation : An Unwritten History (2015)
21. Zordan, V.B., Celly, B., Chiu, B., DiLorenzo, P.C.: Breathe easy. In: Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation - SCA '04. p. 29. ACM Press, New York, New York, USA (2004)