View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Aston Publications Explorer

Speech Breathing in Virtual Humans: An Interactive Model and Empirical Study

Ulysses Bernardet*
School of Engineering and Applied Science
Aston University

Sin-hwa Kang†
Institute for Creative Technologies
University of Southern California

Andrew Feng‡
Institute for Creative Technologies
University of Southern California

Steve DiPaola§
Simon Fraser University

Ari Shapiro ¶
Institute for Creative Technologies
University of Southern California

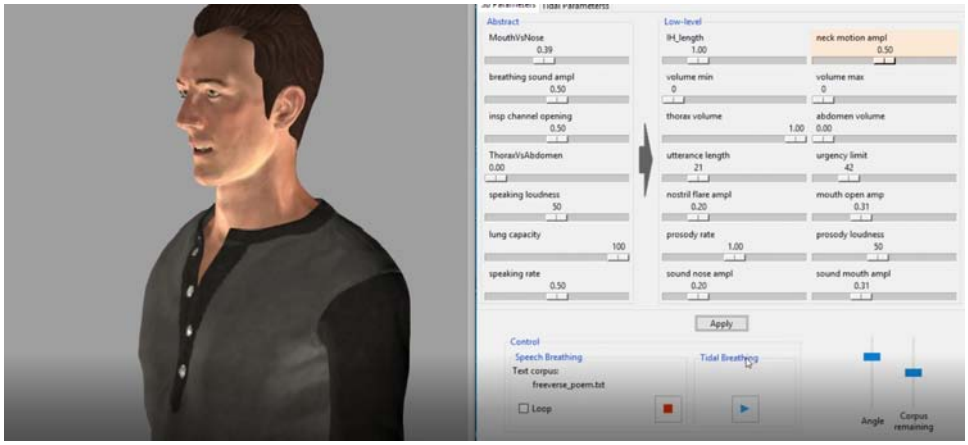

Figure 1: A model for simulating speech breathing. Breaths are added to synthetic speech based on lung capacity, speaking loudness, mouth or nose breathing and other factors.

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).

In addition, we perform a study to determine the effects of including breathing-motivated speech movements, such as head tilts and chest expansions during dialogue on a virtual character, as well as breathing sounds. This study includes speech that is generated both from a text-to-speech engine as well as from recorded voice.

*e-mail:u.bernardet@aston.ac.uk
†e-mail:kang@ict.usc.edu
‡e-mail:feng@ict.usc.edu
§e-mail:sdipaola@sfu.ca
¶e-mail:shapiro@ict.usc.edu

Index Terms: 500 [Human-centered computing]: Virtual reality— [500]: Human-centered computing—Empirical studies in HCI 500 [Computing methodologies]: Procedural animation— [500]: Computing methodologies—Virtual reality

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". We are interested in the impact on the viewer of both the breathing sound (breath intakes), the breathing impact on physiology, such as changes in neck angle and the expansion or contraction of the chest or abdomen, as well as the changes in speech timing that breathing necessitates. From a functional perspective, the respiratory system needs to provide the correct pressure drive to the voice box [13]. Consequently, breathing is implicated in many aspects of speech production [30] such as voice quality, voice onset time, and loudness.

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.

Our contributions include: a model for speech breathing that includes parameterizations of physiological aspects (such as lung volume, mouth versus nose breathing) in combination with speech parameters (such as speaking rate or volume) allowing for a variation in synthetic speech generation. In addition, we perform a user study examining whether the impact of speech breathing sounds or associated speech breathing-related appearance changes the perceptions of the virtual character producing the speech.

2 Related work

2.1 Breathing in virtual human characters

As [28] 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 [22], few virtual character systems actually take breathing into account. In their interactive poker game system, [9] 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 [28, 32]. Recent models are usually based on motion data captured from participants [23, 27]. Modeling work on breathing in conjunction with speaking is sparse, and the work of [12] on the development of infant speech production one of the few published works.

Visual speech systems have focused on the production of lip movements to match the formation of words through phonemes via dominance functions [7], direct representation of diphones or triphones [4, 31] or similar set of decomposed animation curves [26] or machine learning [8, 5, 25]. However, most techniques do not model breathing explicitly, nor model the impact of breathing on the speech systems. Some gesturing methods do include head movements that might relate to breathing, such as nodding the head on the start of a word phrase [16, 17], but such rules are generated as a result of a machine learning process, and not explicitly modeled.

2.2 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 [19]. 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 2 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 [10, 11, 29]. 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 2). In their study [29] 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. We improve on an existing speech breathing model developed in [2] as well conduct an empirical study. 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 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.

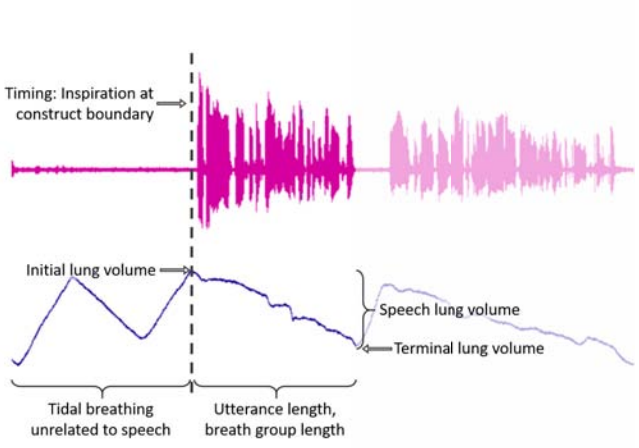

Figure 2: Visualization of the dynamic interplay between breathing and speaking.

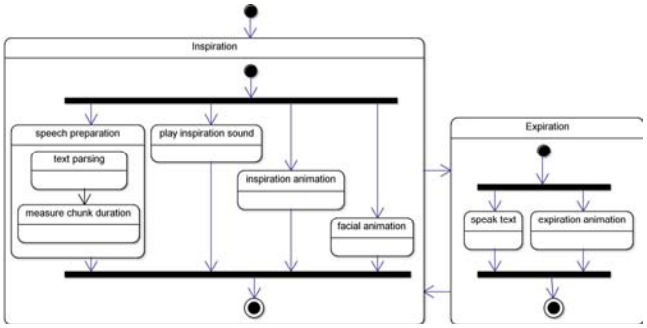

Figure 3: State diagram of the dynamic breathing model

3 Dynamic speech breathing system

3.1 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).

3.2 Control model

At the core of the speech breathing control model stands the oscillation between two fundamental processes: inspiration and expiration (Figure 3).

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 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

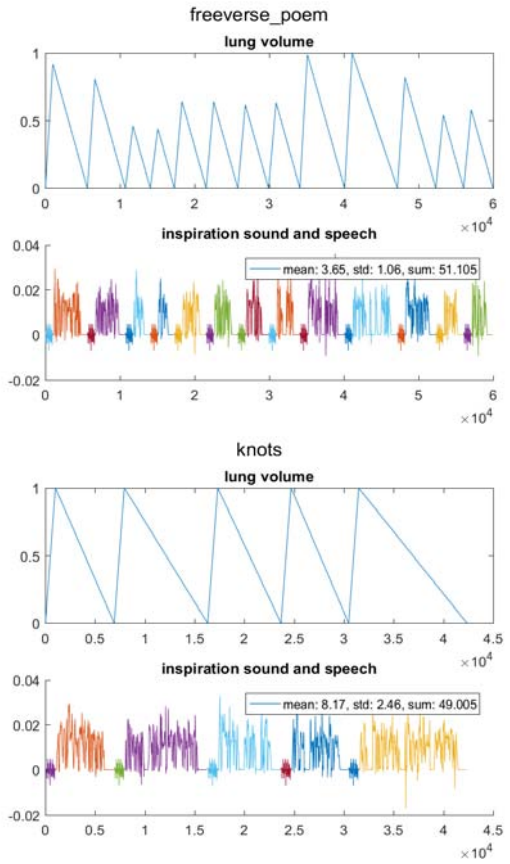

Figure 4: Dynamic behavior of the model for two different texts. The top saw-tooth signal is the driver for the thorax and abdomen shapes, as well as the pitch angle of the head. The lower panel shows the speech and breathing sound outputs.

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:

- 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

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.

3.2.1 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 $amplitude_{neck\ motion}$.

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

The low- and high-level parameters of the model can be controlled in real time using a Graphical User Interface developed using Pygubu [21] and Python's Tkinter module (Figure 5).

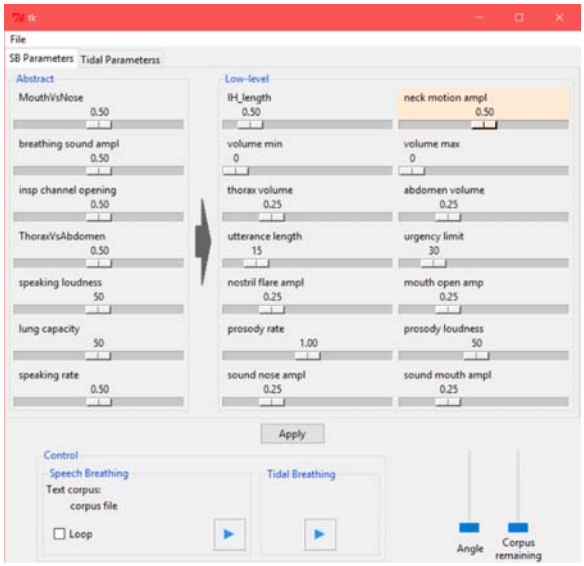

Figure 5: Graphical User Interface for the speech breathing system.

The parameters that can be directly controlled by the user, as well as the low-level parameters that are influenced from those controls are summarized in Table 1.

3.3 Implementation

3.3.1 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 [18]. 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).

Low level	Direct control parameters
$\text{volume}_{\text{abdomen}}$	Maximum volume of the abdomen
$\text{volume}_{\text{thorax}}$	Maximum volume of the thorax
$\text{lenght}_{\text{inspiration}}$	Time to fill the lungs
$\text{amplitude}_{\text{sound mouth}}$	How audible the inspiration through the mouth is
$\text{amplitude}_{\text{sound nose}}$	How audible the inspiration through the nose is
$\text{amplitude}_{\text{mouth open}}$	How much the mouth opens during breathing
$\text{amplitude}_{\text{nostril flare}}$	How much the nostrils flare during breathing
$\text{amplitude}_{\text{neck motion}}$	Head pitch rotation during breathing
$\text{length}_{\text{utterance}}$	Length of the utterance
$\text{urgency}_{\text{limit}}$	Variability of the effective length of the utterance
$\text{prosody}_{\text{loudness}}$	TTS "prosody" loudness
$\text{rate}_{\text{prosody}}$	TTS "prosody" speaking rate
Top level	Parameters that control low-level parameters
ThoraxVsAbdomen	Balance between thoracic and abdominal breating
$\text{capacity}_{\text{lung}}$	Overall lung capacity
MouthVsNose	Balance between breathing though nose vs breathing through mouth
$\text{amplitude}_{\text{breathing sound}}$	How audible the breathing sounds are
$\text{opening}_{\text{insp channels}}$	How much nostrils and mouth are opened during breathing
$\text{speaking}_{\text{loudness}}$	How loud the speech is

Table 1: Users specify the top level parameters (bottom) which in turn influence the low level parameters (top).

.

3.3.2 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, [20]). Concurrently, the controller, via the 'm+m" middleware software [3], sends messages to the SmartBody virtual character system, where the character animation and rendering take place [24]. 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 [6].

4 Method (Empirical Study Design and Results)

4.1 Study Design

One of the key questions in the work presented here is, how relevant is speech breathing behavior for virtual humans. To assess the influence of speech breathing related behavior on the perception of the virtual human, conducted an empirical study where participants rated videos of an animated character. To achieve maximal accuracy of timing and facial expressions we had the character animated by a professional animator (Figure 7).

We used a complete factorial, within subject design with the following factors as independent variables (IV): different speech type (human voice vs. TTS), dialog type (casual tone vs. formal tone,

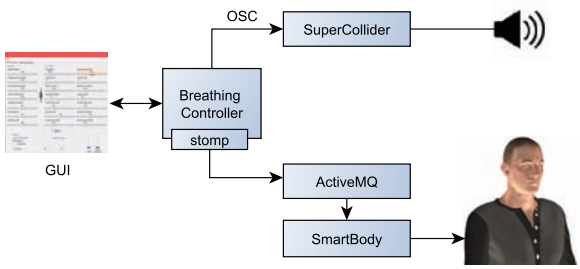

Figure 6: Architecture of the speech breathing control system

[14]), motion type (breathing motion vs. no breathing motion) and breathing sound (audible vs. silent). As dependent variables (DV) the participants on a 7-point Likert scale (1 = "Disagree strongly", 7 = "Agree strongly") rated five impressions of the character, which represent human-like features, depicted in the video: How much they agreed that the character in the video "behaved like a human", "was dynamic", "was attractive", "was trustable", "was likable".

The data was collected using an on-line experiment. A total of 60 participants (age over 18 years old, location: USA) were recruited via the Amazon Mechanical Turk (MTurk) crowdsourcing platform. Prior to the rating task, participants were asked to provide the transcription of a short audio clip. This was to ensure that the sound level was set correctly during the experiment.

4.2 Results

We ran a four-way repeated measures ANOVA using SPSS Statistics to investigate user perception of a virtual character that displayed breathing patterns associated with the four factors described above. Because we had four IV and were interested in understanding which related groups are different at the univariate level, we conducted our analysis on each DV separately using the ANOVA method instead of using the MANOVA method that analyzes data at the multivariate level. The primary purpose of running the four-way repeated measures ANOVA was to understand if there was an interaction between the four factors on DV. We will describe the results of the analysis for each DV below.

4.2.1 How human-like the behavior was

In the results, the epsilon (ε) of 1 that indicates the variances of differences between all possible pairs of groups are equal and thus sphericity is exactly met. Univariate tests show a significant effect of speech type on user perception of the virtual character for "behaving like a human" [F(1, 59) = 84.37, p $<$ .001, η^2 = .59]. Users perceived the character as behaving more like a human when the character spoke using a human voice (Mean = 5.29, SE = .13), compared to a TTS voice (Mean = 3.37, SE = .17). Univariate tests also show that there was a significant effect of dialog type on user

Figure 7: Screen capture of the virtual character used in the evaluation study.

Speech * Dialog * Bsound

Measure:MEASURE_1

Speech	Dialog	Bsound	Mean	Std. Error	Interval Lower Bound	Interval Upper Bound
Human Voice	Casual Tone	Audible	5.258	.158	4.942	5.574
		Silent	5.008	.180	4.648	5.369
	Formal Tone	Audible	5.433	.156	5.122	5.745
		Silent	5.450	.160	5.130	5.770
TTS	Casual Tone	Audible	3.067	.180	2.706	3.427
		Silent	3.292	.185	2.921	3.662
	Formal Tone	Audible	3.633	.176	3.281	3.986
		Silent	3.475	.198	3.079	3.871

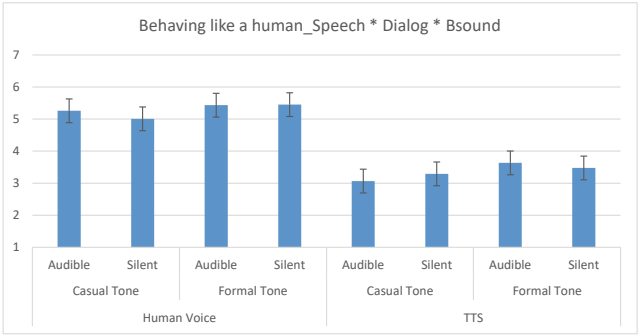

Figure 8: Interaction effect between speech type, dialog type, and breathing sound for "The character behaved like a human"

perception of the virtual character behaving like a human [F(1, 59) = 10.11, p = .002, η^2 = .15]. Users perceived the character as behaving more like a human when the character used a formal tone (Mean = 4.50, SE = .13), in comparison to a casual tone (Mean = 4.16, SE = .11).

However, the results demonstrate that there was an interaction between speech type, dialog type, and breathing sound [F(1, 59) = 7.00, p = .010, η^2 = .11]. Users perceived the character as behaving more like a human when the character spoke with a human voice, audible breathing sound, and casual tone (Mean = 5.26, SE = .16), compared to when the character spoke using a human voice, silent breathing sound, and casual tone (Mean = 5.01, SE = .18). Similar results are shown when the character spoke using a human voice, silent breathing, and formal tone (Mean = 5.45, SE = .16), compared to the character using a human voice, audible breathing sound, and formal tone (Mean = 5.43, SE = .16). Users further perceived the character as behaving more like a human when the character spoke using a TTS voice, silent breathing, and casual tone (Mean = 3.29, SE = .19), compared to the character using a TTS voice, audible breathing, and casual tone (Mean = 3.07, SE = .18). Similar results are shown when the character spoke with a TTS voice, audible breathing, and formal tone (Mean = 3.63, SE = .18), compared to the character using a TTS voice, silent breathing, and formal tone (Mean = 3.48, SE = .20) (Figure 8).

4.2.2 Dynamism of the character

In the results, the epsilon (ε) of 1 that indicates the variances of differences between all possible pairs of groups are equal and thus sphericity is exactly met. Univariate tests show a significant effect of speech type on user perception of the virtual character for "dynamic" [F(1, 59) = 66.02, p $<$ .001, η^2 = .53]. Users perceived the character as a more dynamic one when the character spoke using a human voice (Mean = 5.08, SE = .11), compared to a TTS voice (Mean = 3.55, SE = .16). Univariate tests also show that there was a significant effect of dialog type on user perception of a virtual character for "dynamic" [F(1, 59) = 8.55, p = .005, η^2 = .13]. Users perceived the character as a more dynamic one when the charac-

Dialog * Bmotion

Measure:MEASURE_1

Dialog	Bmotion		Mean	Std. Error	Interval Lower Bound	Interval Upper Bound
	Casual Tone	Breathing Motion	4.229	.134	3.960	4.498
		No Breathing Motion	4.067	.119	3.829	4.304
	Formal Tone	Breathing Motion	4.421	.123	4.175	4.666
		No Breathing Motion	4.546	.115	4.315	4.777

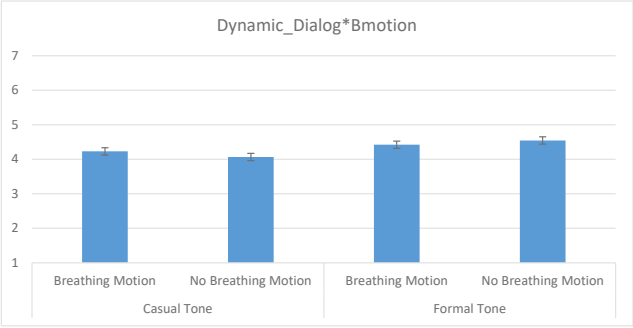

Figure 9: Interaction effect between dialog type and breathing motion for "The character was dynamic"

ter delivered dialog with formal tone (Mean = 4.48, SE = .11), in comparison to casual tone (Mean = 4.15, SE = .12).

However, the results demonstrate that there was an interaction between speech type and dialog type [F(1, 59) = 6.41, p = .014, η^2 = .10]. Users perceived the character as a more dynamic one when the character spoke using a human voice and formal tone (Mean = 5.36, SE = .13), compared to a human voice and casual tone (Mean = 4.80, SE = .16). Similar results are shown when the character spoke using a TTS voice and formal tone (Mean = 3.60, SE = .17), compared to a TTS voice and casual tone (Mean = 3.49, SE = .17). The results also demonstrate that there was an interaction between dialog type and breathing motion [F(1, 59) = 4.18, p = .045, η^2 = .07]. Users perceived the character as a more dynamic one when the character spoke using casual tone with breathing motion (Mean = 4.23, SE = .13), compared to casual tone without breathing motion (Mean = 4.07, SE = .12). Similar results are shown when the character spoke using formal tone without breathing motion (Mean = 4.55, SE = .12), compared to formal tone with breathing motion (Mean = 4.42, SE = .12) (Figure 9).

4.2.3 Attractiveness of the character

In the results, the epsilon (ε) of 1 that indicates the variances of differences between all possible pairs of groups are equal and thus sphericity is exactly met. Univariate tests show a significant effect of speech type on user perception of the virtual character for "attractive" [F(1, 59) = 18.50, p < .001, η^2 = .24]. Users perceived the character as a more attractive one when the character spoke using a human voice (Mean = 4.33, SE = .15), compared to a TTS voice (Mean = 3.93, SE = .16). Univariate tests also show that there was a significant effect of dialog type on user perception of a virtual character for "attractive" [F(1, 59) = 16.00, p < .001, η^2 = .21]. Users perceived the character as a more attractive one when the character delivered dialog with casual tone (Mean = 4.30, SE = .16), compared to formal tone (Mean = 3.97, SE = .15).

The results demonstrate that there were no significant interaction effects between IVs.

4.2.4 How trustable the character was

In the results, the epsilon (ε) of 1 that indicates the variances of differences between all possible pairs of groups are equal and thus sphericity is exactly met. Univariate tests show a significant effect of speech type on user perception of the virtual character for "trustable" [F(1, 59) = 53.87, p < .001, η^2 = .48]. Users perceived the character as a more trustable one when the character spoke using a human voice (Mean = 4.91, SE = .14), compared to a TTS voice (Mean = 3.79, SE = .15). Univariate tests also show that there was a significant effect of dialog type on user perception of a virtual character for "trustable" [F(1, 59) = 11.93, p = .001, η^2 = .17]. Users perceived the character as a more trustable one when the character delivered dialog with formal tone (Mean = 4.54, SE = .14), compared to casual tone (Mean = 4.16, SE = .13).

However, the results also demonstrate that there was an interaction between speech type and breathing motion [F(1, 59) = 4.04, p = .049, η^2 = .06]. Users perceived the character as a more trustable one when the character spoke using a human voice without breathing motion (Mean = 4.98, SE = .14), compared to a human voice with breathing motion (Mean = 4.83, SE = .15). Similar results are shown when the character spoke using a TTS voice with breathing motion (Mean = 3.82, SE = .14), compared to a TTS voice without breathing motion (Mean = 3.75, SE = .16) (Figure 10). The results further demonstrate that there was an interaction between speech type, dialog type, and breathing sound [F(1, 59) = 5.89, p = .018, η^2 = .09]. Users perceived the character as a more trustable one when the character spoke using a human voice and casual tone with audible breathing (Mean = 4.81, SE = .16) than a human voice and casual tone with silent breathing (Mean = 4.60, SE = .18). Similar results are shown when the character spoke using a human voice and formal tone with silent breathing (Mean = 5.13, SE = .16) than a human voice and formal tone with audible breathing (Mean = 5.11, SE = .16). Users further perceived the character as a more trustable one when the character spoke using a TTS voice and casual tone with silent breathing (Mean = 3.71, SE = .16) than a TTS voice and casual tone with audible breathing (Mean = 3.52, SE = .16). Similar results are shown when the character spoke using a TTS voice and formal tone with audible breathing (Mean = 4.04, SE = .16) than a TTS voice and formal tone with silent breathing (Mean = 3.88, SE = .18) (Figure 11).

4.2.5 Likability of the character

In the results, the epsilon (ε) of 1 that indicates the variances of differences between all possible pairs of groups are equal and thus sphericity is exactly met. Univariate tests show a significant effect of speech type on user perception of the virtual character for "likable" [F(1, 59) = 65.83, p < .001, η^2 = .53]. Users perceived the character as a more likable one when the character spoke with a human voice (Mean = 4.94, SE = .14), compared to a TTS voice (Mean = 3.71, SE = .15). Univariate tests also show that there was a significant effect of dialog type on user perception of a virtual character for "likable" [F(1, 59) = 4.81, p = .03, η^2 = .08]. Users perceived the character as a more likable one when the character delivered dialog with formal tone (Mean = 4.42, SE =.13), compared casual tone (Mean = 4.24, SE = .13).

However, the results demonstrate that there was an interaction between speech type and dialog type [F(1, 59) = 18.24, p < .001, η^2 = .24]. Users perceived the character as a more likable one when the character spoke using a human voice and formal tone (Mean = 5.26, SE = .16), in comparison to a human voice and casual tone (Mean = 4.63, SE = .17). Similar results are shown when when the character spoke using a TTS voice and casual tone (Mean = 3.85, SE = .16), compared to a TTS voice and formal tone (Mean = 3.58, SE = .15).

Overall, users gave a higher rating when the character used a human voice for all of the 5 DVs, rather than a computer-generated

Speech * Bmotion

Measure:MEASURE_1

Speech	Bmotion	Mean	Std. Error	Interval Lower Bound	Interval Upper Bound
Human Voice	Breathing Motion	4.833	.147	4.539	5.127
	No Breathing Motion	4.983	.136	4.712	5.255
TTS	Breathing Motion	3.821	.143	3.535	4.106
	No Breathing Motion	3.754	.161	3.432	4.076

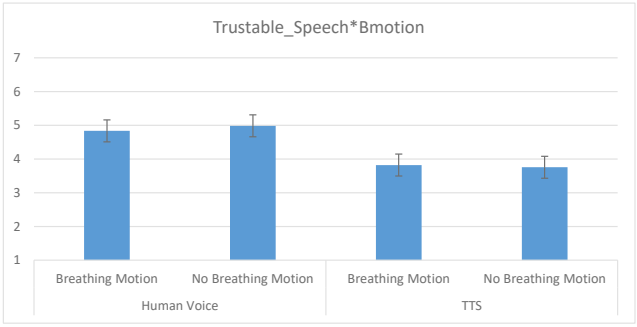

Figure 10: Interaction effect between speech type and breathing motion for "The character was trustable"

TTS voice. This is not a big surprise. Further, participants gave a higher rating when the character spoke with formal tone, rather than casual tone, for the 4 DVs. The voice in the casual tone dialog may have seemed less sophisticated than the voice in the formal tone dialog, which employed a professional weather forecaster. Thus, this does not seem to be a big surprise either. However, there are significant interaction effects between some of the IVs. Further understanding of these interaction effects described below may better answer our research questions.

4.3 Implications of interaction effects

4.3.1 Interaction between speech type and dialog type

There was a significant interaction effect for speech type (human or TTS) and dialog type (formal or casual) for only the dynamic and likeable measures. The character using formal tone with either speech type, was regarded as more dynamic and likable features than the character using casual tone with either speech type. These findings partially resonate with higher ratings for the main effects of the character using formal tone across the 4 DVs.

4.3.2 Interaction between speech type and breathing motion

There was a significant interaction effect for speech type (human or TTS) and breathing motion (with or without motion) for only the trustable measure. Participants perceived the character presenting either a human voice without breathing motion or a TTS voice with breathing motion as a more trustable one than the other combinations of the variables. These findings indicate participants perceive the character as a trustable one when it speaks using a human voice as the voice itself is realistic. While at the same time, they expect to see breathing motion for the character with the computer-generated voice, since they want it to seem more human in order to trust it.

4.3.3 Interaction between dialog type and breathing motion

There was a significant interaction effect for dialog type (formal or casual) and breathing motion (with or without motion) for only

Speech * Dialog * Bsound

Measure:MEASURE_1

Speech	Dialog	Bsound	Mean	Std. Error	Interval Lower Bound	Interval Upper Bound
Human Voice	Casual Tone	Audible	4.808	.161	4.487	5.130
		Silent	4.592	.184	4.224	4.960
	Fomal Tone	Audible	5.108	.160	4.787	5.429
		Silent	5.125	.159	4.807	5.443
TTS	Casual Tone	Audible	3.517	.155	3.207	3.827
		Silent	3.708	.156	3.395	4.021
	Fomal Tone	Audible	4.042	.160	3.722	4.361
		Silent	3.883	.183	3.517	4.250

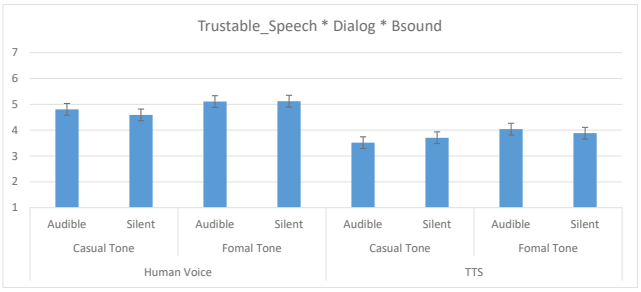

Figure 11: Interaction effect between speech type, dialog type, and breathing sound for "The character was trustable"

the dynamic measures. Participants perceived the character delivering dialog using either casual tone with breathing motion or formal tone without breathing motion as more dynamic one than the dialogues with other combinations. The dialog with casual tone was captured from a radio show addressing advertising revenues, while the dialog with formal tone was captured from a TV show delivering weather forecasts. The casual tone dialog was conveyed with enthusiasm, thus participants might have felt the character speaking the dialog accompanied by breathing motion was more realistic. The formal tone dialog did not present any emotional signal, thus participants might have perceived the character speaking the dialog without breathing motion as more realistic since the computer generated breathing motion could have seemed incompatible with the matter-of-fact tone.

4.3.4 Interaction between speech type, dialog type, and breathing sound

There was a significant interaction effect for speech type (human or TTS), dialog type (formal or casual), and breathing sound (silent or audible) for only the behaving like a human and trustable measures. Participants perceived the character as behaving more like a human and a more trustable one when it spoke using either a human voice, casual tone, and audible breathing, or a human voice, formal tone, and silent breathing. This perhaps indicates that participants might have felt the character speaking with casual tone and breathing sound was more congruent and therefore realistic as the dialog with casual tone conveyed enthusiasm. The dialog with formal tone did not convey a strong emotional signal, thus participants might have perceived the character speaking the dialog without breathing sound as more professional and the computer-generated breathing sound would have seemed out of place.

Unlike characters with a human voice, participants perceived the TTS voice using character as behaving more like a human and a more trustable one when it delivered TTS dialog using either the casual tone without breathing sound or the formal tone with audible breathing. This indicates that participants might have felt the character speaking dialog using the formal tone and the breathing sound was more realistic as the breathing helped portray the TTS using character as more human. The dialog with casual tone and the breathing sound may have been perceived as too incongruent

with the TTS voice. Perhaps it seemed that it was trying too hard to sound human.

5 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.

The results of our study suggest to design a human-like virtual character by implanting breathing motion in the character when it delivers casual and enthusiastic tone dialog. It is also proposed to use a TTS voice with breathing motion to create the human-like character. It is further suggested to design the human-like character using either a human voice with breathing sound or a TTS voice without breathing sound when the character delivers casual tone dialog, while using either a human voice without breathing sound or a TTS voice along with audible breathing when the character delivers formal tone dialog.

5.1 Limitations

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 [15] might be important for realism.

5.2 Future steps

Future steps include the improvement of the breathing animation to a state of the art implementation as presented e.g. in [27]. 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).

This effort was supported by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

[1] http://www.audacity.audio/.

[2] U. Bernardet, S.-h. Kang, A. Feng, S. DiPaola, and A. Shapiro. A dynamic speech breathing system for virtual characters. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, editors, *Intelligent Virtual Agents*, pages 43–52, Cham, 2017. Springer International Publishing.

[3] U. Bernardet, T. Schiphorst, D. Adhia, N. Jaffe, J. Wang, M. Nixon, O. Alemi, J. Phillips, S. DiPaola, and P. Pasquier. 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*, pages 1–21, New York, New York, USA, 2016. ACM Press.

[4] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, pages 353–360, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[5] Y. Cao, P. Faloutsos, and F. Pighin. Unsupervised learning for speech motion editing. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 225–231. Eurographics Association, 2003.

[6] http://www.cereproc.com/.

[7] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, 1993.

[8] Z. Deng, P.-Y. Chiang, P. Fox, and U. Neumann. Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, I3D '06, pages 43–48, New York, NY, USA, 2006. ACM.

[9] P. Gebhard, M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler, and O. Türk. IDEAS4Games: Building expressive virtual characters for computer games. In *Lecture Notes in Computer Science*, volume 5208 LNAI, pages 426–440, 2008.

[10] A. Henderson, F. Goldman-Eisler, and A. Skarbek. Temporal Patterns of Cognitive Activity and Breath Control in Speech. *Language and Speech*, 8(4):236–242, 1965.

[11] T. J. Hixon, M. D. Goldman, and J. Mead. 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.

[12] I. S. Howard and P. Messum. Modeling Motor Pattern Generation in the Development of Infant Speech Production. *8th International Seminar on Speech Production*, pages 165–168, 2008.

[13] J. E. Huber and E. T. Stathopoulos. Speech Breathing Across the Life Span and in Disease. *The Handbook of Speech Production*, pages 11–33, 2015.

[14] Kate Moran. The Four Dimensions of Tone of Voice, 2016.

[15] D. R. Ladd. Declination: a review and some hypotheses. *Phonology*, 1(1):53–74, 1984.

[16] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, editors, *Intelligent Virtual Agents*, pages 243–255, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[17] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, pages 25–35, New York, NY, USA, 2013. ACM.

[18] J. McCartney. Rethinking the Computer Music Language: SuperCollider. *Computer Music Journal*, 26(4):61–68, 12 2002.

[19] D. H. McFarland and A. Smith. Effects of vocal task and respiratory phase on prephonatory chest wall movements. *Journal of speech and hearing research*, 35(5):971–82, 10 1992.

[20] http://opensoundcontrol.org/.

[21] https://github.com/alejandroautalan/pygubu.

[22] J. Rickel, U. S. C. Information, E. André, N. Badler, and J. Cassell. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, 2002.

[23] B. Sanders, P. Dilorenzo, V. Zordan, and D. Bakal. Toward Anatomical Simulation for Breath Training in Mind/Body Medicine. In N. Magnenat-Thalmann, J. J. Zhang, and D. D. Feng, editors, *Recent Advances in the 3D Physiological Human*. Springer, 2009.

[24] A. Shapiro. Building a character animation system. In *Motion in Games*, pages 98–109, 2011.

[25] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *ACM Trans. Graph.*, 36(4):93:1–93:11, July 2017.

[26] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews. Dynamic units of visual speech. In *Proceedings of the 2012 ACM SIGGRAPH/Eurographics symposium on Computer animation*, jul 2012.

[27] A. Tsoli, N. Mahmood, and M. J. Black. Breathing life into shape. *ACM Transactions on Graphics*, 33(4):1–11, 7 2014.

[28] R. C. Veltkamp and B. Piest. A Physiological Torso Model for Realistic Breathing Simulation. In *Proceeding 3DPH'09 Proceedings of the 2009 international conference on Modelling the Physiological Human*, pages 84–94, 2009.

[29] A. L. Winkworth, P. J. Davis, R. D. Adams, and E. Ellis. Breathing patterns during spontaneous speech. *Journal of speech and hearing research*, 38(1):124–144, 2 1995.

[30] M. Włodarczak, M. Heldner, and J. Edlund. Breathing in Conversation : An Unwritten History, 2015.

[31] Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, MIG '13, pages 109:131–109:140, New York, NY, USA, 2013. ACM.

[32] V. B. Zordan, B. Celly, B. Chiu, and P. C. DiLorenzo. Breathe easy. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation - SCA '04*, page 29, New York, New York, USA, 2004. ACM Press.