Audio Analysis of Human/Virtual-Human Interaction

Harold Rodriguez[1], Diane Beck[1], David Lind[2], and Benjamin Lok[1]

[1] University of Florida, Gainesville FL 32607, USA,
drharold@ufl.edu, lok@cise.ufl.edu, beck@cop.ufl.edu,
WWW home page: http://www.cise.ufl.edu/research/vegroup/
[2] Medical College of Georgia, Augusta GA 30912, USA,
dlind@mail.mcg.edu

Abstract. The audio of the spoken dialogue between a human and a virtual human (VH) is analyzed to explore the impact of H-VH interaction. The goal is to determine if conversing with a VH can elicit detectable and systematic vocal changes. To study this topic, we examined the H-VH scenario of pharmacy students speaking with immersive VHs playing the role of patients. The audio analysis focused on the students' reaction to scripted empathetic challenges designed to generate an unrehearsed affective response from the human. The responses were analyzed with software developed to analyze vocal patterns during H-VH conversations. The analysis discovered vocal changes that were consistent across participants groups and correlated with known H-H conversation patterns.

Keywords: Conversational agents, speech visualization, digital signal processing

1 Introduction

Recently, significant work has been applied towards providing opportunities to practice human-human interactions through interactions with a virtual human (VH). These VHs play the role of conversational agents and have been used in military [1], medical [2], and entertainment [3] scenarios to train soldiers, detect empathy, and practice communication skills. In these systems, there has been a consistent effort to increase the communication cues that are tracked by the VH simulation to increase interaction fidelity.

In this paper, we focus on the information contained within the H-VH audio stream. The goal of this work is to determine if the audio stream of a person speaking with a VH contains prosodic information that can 1) quantify the impact of the VH experience and 2) be used to improve the VH experience. That is, can VHs elicit detectable and systematic vocal changes in humans? Using a digital signal processing approach, critical moments in a H-VH conversation are analyzed to measure the human response to VHs.

First, this work presents a signal processing approach to H-VH audio analysis. The approach focuses on the application of frequency-spectrum audio analysis and visualization techniques. It describes the most applicable metrics to quantifying the user's tone while conversing with a VH. Second, a user study designed to generate an unrehearsed affective response from the human is used to investigate the efficacy of the approach. Finally, this paper presents directions on applying the detectable vocal changes as an input into VH simulations, as signs of user attention and as a valuable teaching tool to help poor students remediate their communication skills.

Related Work. Significant work on vocal analysis and emotion detection has been conducted by J. Bachorowski, K. Scherer, R. Cowie, and others. Studies have shown the value of extracting speech or acoustic *parameters* [4, 5], features that describe a person's speech, as opposed to focusing on lexical and grammatical measures. In fact, these acoustic parameters are language-independent [6, 7] and convey subtle information to the listener. For example, lowered tone and a softer voice may indicate sadness without having to verbally express this sentiment. Hence, some preliminary emotion detection has been attempted using these attributes, based on prosodic training algorithms [8], Bayesian inference [9], and promotion/demotion schemes [10].

It is useful to set aside strict emotion detection and instead focus on understanding the basic acoustic properties that provide the best heuristics for researchers in this field. There is only a weak link between speech streams and their signature traits because a good "Rosetta stone"to decode the traits has yet to be developed [11]. However, a select few have shown to be significant: fundamental frequency (F0), speech rate, and amplitude [12]. These three main measures, a few novel ones, and some from [13] are explored within a H-VH context.

2 A Signal Processing Approach to H-VH Audio Analysis

Overview. Spoken dialogue, natural language, and embodied conversational agents systems (among many others) all aim to derive higher level meaning from the content of the words presented by the user. This work aims to supplement the content information by quantifying global and local *frequency* and *temporal* metrics of the user's audio.

In this section, we will describe an approach for analyzing user speech as well as software that was developed to support the unique requirements of H-VH interaction analysis. In our user studies (Sect. 3), each student had a conversation with a VH for about 13 minutes. The audio was extracted from WMV video file recordings of the students (32 kHz, 16 bit, 32 kbps) and the audio was loaded into CFMetrix, an in-house developed audio analysis and visualization program that processes the data. Coupled with user interaction, the system highlights significant moments in the audio stream.

Metrics. In processing the audio signal of a H-VH interaction, there are dozens of potential metrics to compute and analyze. However, metrics which have been shown to have a large impact on H-H conversations [12, 4, 5] were implemented and grouped by domain (frequency or temporal).

Frequency metrics focus on the frequency domain characterization of the user's audio stream. While the primary frequency metric is the F0 (fundamental frequency) of the user's audio, other frequency metrics can be derived:

- *F0 Mean* - the average fundamental frequency (main pitch) of the signal over some time. F0 is obtained by performing an FFT of the audio stream and finding the first major peak in the signal (typically around 120 Hz for men and 220 Hz for women).
- *F0 Variance* - the statistical dispersion of the fundamental frequency over the support time. Defines amount of inflection in an utterance.
- *Frequency Range* - the largest range of frequencies above a given power threshold. Provides the size of most speech energy.
- *Frequency Histogram* - a running sum of waveform amplitudes over all available frequencies. Conveniently exaggerates the locations of user speech and ambient noise.

Temporal metrics focus on the "loudness"of speech and speech timing. In addition to the audio's signal amplitude, RMS amplitude, power mean, and power variance, the following calculations are useful:

- *Power Level* - $10 * log_{10}(Power/10pW)$. Relatable loudness (units in dB).
- *Power Range* - RMS amplitude of a superposition of FFTs over all time. Typical loudness of a user.
- *Speaking Contour* - represented as a graph. When someone is speaking, a rising slope is shown, but as soon as the voice ceases, it drastically declines.
- *Signal-to-noise ratio (SNR)* - defined as $10 * log10(Power/Power_{min})$, where $Power_{min}$ is a user-defined minimum power level. Global SNR provides information on the viability of ambient noise filtering, whereas local SNR justifies the inclusion or exclusion of an utterance for analysis.
- *Rate of Speech* - divides the number of words (number of times the power of the spectrum is above $Power_{min}$) by the sound duration (in minutes).

Filters and Visualizations. The audio stream of a human talking to a VH has unique properties that assist in analysis. Specifically, in most H-VH interactions 1) usually only one conversationalist is speaking, 2) audio recordings of the user are separable since they have a high SNR, and 3) the user is speaking into a speech recognition system (or is under that assumption) and thus tends to enunciate more than usual.

Nevertheless, several averaging, normalization, and bandpass filters were created and used to aid signal "clean-up". The visualizations provide instant feedback on the effect of these adjustable filters. Furthermore, the researcher can interact with visualizations to identify the sound level under which no meaningful signal lies. This made for a very easy, intuitive, and effective model for facilitating metric calculations.

Metrics (e.g. F0) were rendered in both 2D and 3D modes (Fig. 1). In the 2D visualization, frequency vs. amplitude was visualized in real-time. Multiple audio streams can be easily compared using such a visualization: A "difference

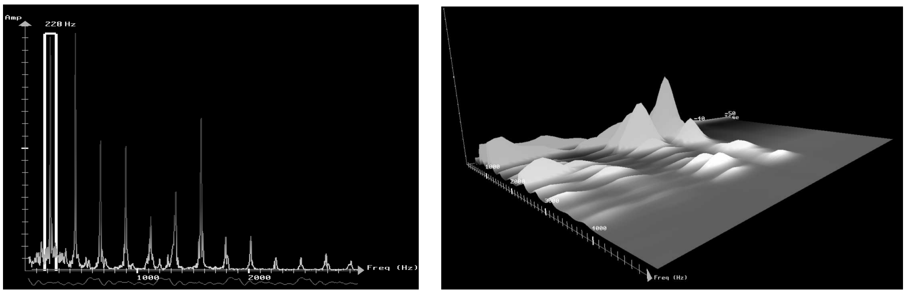

Fig. 1. Left: A white box whose width is proportional to the F0 Variance is centered on the F0 of a user saying "Uhh." **Right**: 3D FFT visualization of a sound stream (*horiz., vert.*) over time (*depth*) with moving average filtering enabled (for smoother heightmap).

spectrum"is generated by analyzing two streams at once. The differences in their FFT spectra, as well as differences in metrics, can be color-coded and displayed simultaneously. Thus, 2D frequency visualizations enable an efficient method to visually explore audial metrics and the impact of filter parameters.

The 3DFFT visualization presents the Fourier transform of the sound stream over an extended time (e.g. 10 seconds), allowing a user to not only identify frequency-related metrics at a certain time (2D) but trends and metric interactions over time. Consequently, this was one method used to identify metrics to explore in the user study.

3 User Study

Motivation. Given a system that can analyze the audio from a H-VH interaction, we examined 27 audio streams of pharmacists talking to a VH playing the role of a patient. The study had the following three goals:

1. Verify the metrics most applicable to processing a conversation with a VH.
2. Identify if audio analysis could characterize a H-VH interaction.
3. Understand the subconscious, affective impact of speaking with a VH.

To magnify the empathetic response, the audio analysis focused on "empathetic challenges". An empathetic challenge is VH-initiated dialogue meant to generate an unrehearsed affective response from the human (e.g. "I'm scared. Could this be cancer?"). Empathetic challenges were chosen as they tend to elicit the strongest, most genuine responses in previous studies [2] and were good indicators of user involvement [14].

Test-bed platform: The Interpersonal Simulator (IPS). The Interpersonal Simulator (IPS) is a VH experience used to train interpersonal skills [15]. The IPS combines VR tracking, natural speech processing, and immersive high-fidelity VHs to produce an engaging VH experience. It is constantly evolving through extensive user testing with over 412 medical, nursing, physician assistant, and pharmacy students and professionals. The virtual experience has been compared to H-H interaction [14], demonstrated construct validity for the evaluation of interpersonal skills [16], and applied to scenarios involving empathy [17, 2], sexual history [18], and breast cancer [19].

Application: Pharmacy Patient Interview. Interpersonal skills have been shown to improve patient care and reduce litigation costs, making them critical for health professionals [20, 21]. For this study, the researchers collaborated with a pharmacy college to train practicing pharmacists with H-VH interactions. Instead of training with human actors pretending to have a disease, the pharmacists interacted with a VH. The scenario was developed based on a standard pharmaceutical peptic ulcer disease case found in [22]. The VH simulates a 35-year old Caucasian male patient complaining about increasing pain in his abdominal region. A special response occurred during the scenario denoted as an *empathetic moment.*

About seven minutes into the interview, the VH relates to the participant that "my dad died of cancer"and asks, "Could this be cancer?"This moment was designed to evoke sadness and an empathetic response from the participant. Pharmacists are trained to handle this situation delicately and professionally, expressing empathy and providing reassurance that the medical team will do everything they can to find out what is wrong.

Thirty-nine pharmacy students (12 men, 27 women) participated in the study. Of these, audio for 8 men and 19 women was recorded (n=27). The goal was for the pharmacists to 1) comfort the concerned patient and 2) identify the cause of the pain. A Wizard-of-Oz (WoZ) interface was used to prevent (poor) speech recognition performance from detrimentally influencing user tone. Audio responses were recorded using a wireless microphone attached to a baseball cap used for head-tracking (Fig. 2), and all audio was analyzed after the experiment had concluded.

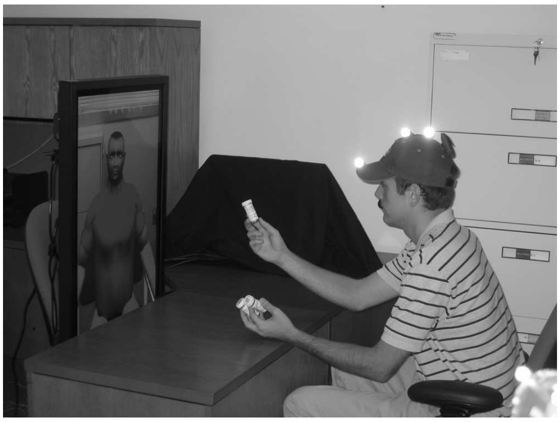

Fig. 2. A user interacts with the VH using natural speech.

Data Analysis. The audio files were analyzed at two levels of granularity, global and local. The global metrics were computed by splitting the audio file at the empathetic moment (≈7 minutes into the conversation). The two parts are labeled Global_{before} and Global_{after}. For local metrics, one sentence before and one sentence after the empathetic challenge were analyzed, labeled Local_{before} and Local_{after}. So for each users speech, each of the metrics in Sect. 2 was computed four times. The four comparisons made are:
1) Local_{before} vs. Local_{after} the immediate impact of the empathetic moment
2) Global_{before} vs. Global_{after} the impact of conversing with a VH for ≈7 min.
3) Global_{before} vs. Local_{after} the startle response of a VH challenge
4) (Local_{after} - Local_{before}) vs. (Global_{after} - Global_{before}) - compares the measuring time-span's effect on metrics

Results. *Vocal changes were detected and measured at the empathetic moment.* A "downward"trend in user pitch and volume arose, which could signal an empathetic response and highlight the VHs affective capabilities.

Paired two-tailed Student T-Tests were performed on the data set. As seen in Table 1, the F0 variance, range, signal amplitude, RMS amplitude, and power level of the audio was significantly different ($p<0.03$) after the empathetic moment.

Table 1. Local user vocal changes near the empathetic moment.

Metric	Change	%	σ	P-value
F0 Mean (Hz)	-8	-4	2.02	.002
Freq. Range (Hz)	-120	-11	27.5	.022
RMS Amplitude	-266	-17	13.8	.006
Power Level (dB)	-1.39	-2	.082	.024

Gender has an impact on tonal change elicited by the VH. F0 mean and RMS amplitude were significantly lower ($p<0.03$) immediately after the challenge for male participants. On top of that, women also showed a significant change in frequency range and power level (Table 2, 3). F0 measurements tend to be reliable indicators of gender [23], but an affective situation such as an empathetic challenge could invalidate this assumption, as F0 is rather unstable here.

Analyzing the local *effect of the empathetic challenge is a good indicator of the entire conversation.* (Local_{after} - Local_{before}) significantly correlated with (Global_{after} vs. Global_{before}). P-values were high ($p>.5$) meaning that there is no statistical difference attributed to changing granularity when considering differences. Hence, a person's reaction to the empathetic challenge permeated across to the other side of the global arena as the user was compelled to extend the VH experience beyond that of a mere chatbot.

Table 2. Local vocal change for men.

Metric	Change	%	σ	P-value
F0 Mean	-7	-5	.787	.018
RMS Amplitude	-424	-30	20.6	.049

Table 3. Local vocal change for women.

Metric	Change	%	σ	P-value
F0 Mean	-8	-4	.589	.014
Freq. Range	-132	-11	-29.2	.024
RMS Amplitude	-199	-12	10.6	.038
Power Level	-1.1	-2	.062	.051

The trend to permeate on until the end of the interaction was realized when manually comparing the raw percentage differences: for all statistically significant metrics (and even some which were not shown to be significant), the local vs. global values were all "in step"(i.e. if one dropped, the other dropped).

4 Conclusions and Future

This paper proposed a signal analysis approach to exploring the audio of a H-VH interaction. The user's vocal metrics changed in a manner similar to those in H-H interactions [5, 24], indicating a subconscious response to the VH. Hence, a strong case is made for having a system capable of tracking and responding to audio changes, thus enhancing the use of VHs as a valuable teaching tool. Particularly *real-time* analysis of the user's speech should be explored and fed as an input to VH simulations for judging attention and engagement.

Further social experiments can be conducted making vocal comparisons of the same user in various scenarios (or different users in the same scenario) to identify the inherent differences of the interaction. For example, this would enable systems to identify if business professionals speak in a different tone with VHs representing clients with different racial, gender, and weight backgrounds. Most of all, VH researchers would benefit from employing the use of metrics, methods, and correlations of these auditory research findings to reinforce and support their affection results. It is through this integration of validated metrics and findings that the VHs' effects can best be understood.

References

1. Swartout, W., e.a.: Toward the holodeck: Integrating graphics, sound, character, and story. In: Proceedings of the 5th International Conference on Autonomous Agents, ACM Press, New York (2001) 409–416
2. A. Deladisma, M. Cohen, e.a.: Do medical students respond empathetically to a virtual patient? American Journal of Surgery **193**(2) (2007) 756–760
3. Dow, S., e.a.: Presence and engagement in an interactive drama. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, ACM Press, New York (2007) 409–416
4. Scherer, K.R.: Vocal affect expression: A review and model for future research. Psychological Bulletin **99**(2) (1986) 143–165

5. Cowie, R., e.a.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **1** (2001) 32–80
6. Scherer, K.R., e.a.: Emotion inferences from vocal expression correlate across languages and cultures. Journal of Cross-Cultural Psychology **32**(1) (2001) 76–92
7. Campbell, N.: Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation. In: ICSLP Proceedings. (2004) 881–884
8. Polzin, T.S., Waibel, A.: Emotion-sensitive human-computer interfaces. In: Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. (September 2000) 201–206
9. McGilloway, S., e.a.: Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. (September 2000) 207–212
10. Rodriguez, H.: Detecting emotion in the human voice: a model and software implementation using the FMOD sound API. Unpublished. (2005)
11. Russell, J.A., Bachorowski, J.A., Fernandez-Dols, J.M.: Facial and vocal expressions of emotion. In: Annual Review of Psychology. (February 2003) 329–349
12. Owren, M.J., Bachorowski, J.A. In: Measuring Emotion-Related Vocal Acoustics. Oxford University Press Series in Affective Sciences (2004)
13. Scherer, K., Banse, R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology **70**(3) (1996) 614–636
14. A. Raij, K. Johnsen, e.a.: Comparing interpersonal interactions with a virtual human to those with a real human. IEEE Transactions on Visualization and Computer Graphics **13**(3) (2007) 443–457
15. K. Johnsen, R. Dickerson, e.a.: Evolving an immersive medical communication skills trainer. Presence: Teleoperators and Virtual Environments **15** (2006) 33–46
16. K. Johnsen, A. Raij, e.a.: The validity of a virtual human system for interpersonal skills education. ACM SIGCHI (2007)
17. M. Cohen, A. Stevens, e.a.: Do health professions students respond empathetically to a virtual patient? Presented at Southern Group on Educational Affairs (2006)
18. A. Deladisma, D. Mack, e.a.: Virtual patients reduce anxiety and enhance learning when teaching medical student sexual-history taking skills. Association for Surgical Education 2007 Surgical Education Week **193**(2) (2007) 756–760
19. A. Kotranza, D. Lind, C.P., Lok, B.: Virtual human + tangible interface = mixed reality human. an initial exploration with a virtual breast exam patient. In: IEEE Virtual Reality 2008. (2008) 99–106
20. C. Vincent, M.Y., Phillips, A.: Why do people sue doctors? a study of patients and relatives taking legal action. Obstetrical and Gynecological Survey **50** (1995) 103–105
21. F.D. Duffy, G.H. Gordon, e.a.: Assessing competence in communication and interpersonal skills: The kalamazoo ii report. Academic Medicine **79** (2004) 495–507
22. Schwinghammer, T.L.: Pharmacotherapy Casebook: A Patient-Focused Approach. McGraw-Hill Medical (2005)
23. Childers, D., Wu, K.: Gender recognition from speech. part ii: Fine analysis. J. Acoust. Soc. Amer. **90** (1991) 18411856
24. Murray, I., Arnott, J.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Amer. **93**(2) (1993) 1097–1108