Logopedics Phoniatrics Vocology, 2013; Early Online: 1–11

informa
healthcare

ORIGINAL ARTICLE

Exploring expressivity and emotion with artificial voice and speech technologies

SANDRA PAULETTO[1], BRUCE BALENTINE[2], CHRIS PIDCOCK[3], KEVIN JONES[4], LEONARDO BOTTACI[5], MARIA ARETOULAKI[6], JEZ WELLS[7], DARREN P. MUNDY[8] & JAMES BALENTINE[9]

[1]*Department of Theatre, Film and Television, University of York, UK,* [2]*Enterprise Integration Group, Zürich, Switzerland,* [3]*CereProc Ltd, Edinburgh, UK,* [4]*Composer and Sound Artist, USA* [5]*Computer Science, University of Hull, UK,* [6]*DialogCONNECTION Ltd, Manchester, UK,* [7]*Department of Music, University of York, UK,* [8]*School of Arts and New Media, University of Hull, UK, and* [9]*Department of Music, University of Texas at San Antonio, Texas, USA*

Abstract
Emotion in audio-voice signals, as synthesized by text-to-speech (TTS) technologies, was investigated to formulate a theory of expression for user interface design. Emotional parameters were specified with markup tags, and the resulting audio was further modulated with post-processing techniques. Software was then developed to link a selected TTS synthesizer with an automatic speech recognition (ASR) engine, producing a chatbot that could speak and listen. Using these two artificial voice subsystems, investigators explored both artistic and psychological implications of artificial speech emotion. Goals of the investigation were interdisciplinary, with interest in musical composition, augmentative and alternative communication (AAC), commercial voice announcement applications, human–computer interaction (HCI), and artificial intelligence (AI). The work-in-progress points towards an emerging interdisciplinary ontology for artificial voices. As one study output, HCI tools are proposed for future collaboration.

Key words: *Artificial emotion, chatbot, conversational kiosk, speech synthesis, TTS, voice synthesis*

Introduction

This report describes work-in-progress by the Voice Expressivity and Emotion Group (VEEG), a working group of the Creative Speech Technology (CreST) network. The initial goal of the network was to plan, design, build, and then present in a public roadshow 'some sort of artefact relevant to speech technology' (1). The public roadshow event, completed in January 2013, constitutes the output of this first phase of research.

Emotion, expression, and communication

Language and meaning play an important role in the portrayal of emotions in a spoken text. However, it is through the sound of the voice that the emotional content is confirmed and emphasized. Experiencing emotions produces physiological changes that affect the way we speak, and 'even slight changes in physiological regulation will produce variations in the acoustic pattern of the speech waveform' (2, p. 240).

Various studies have analysed the characteristics of acoustic parameters in emotional speech signals. They are usually based on the *standard content paradigm* (3, in 8) in which subjects speak the same text with different emotional intentions and then variations in acoustic parameters are measured. The assumption in this method is that, because the text remains the same, the variations in acoustic parameters are the only conduit for the different emotional intentions.

This and other studies (4–9) have found that the main acoustic contributors to the display of emotional speech are fundamental frequency, the intensity or loudness of the voice, the energy distribution in the signal spectrum (e.g. stronger upper partials), and lastly speech rate, which includes the speed at which words are spoken

Correspondence: Dr Sandra Pauletto, Director of MA/MSc in Postproduction with Sound Design, Department of Theatre, Film and Television, The University of York, East Campus, Baird Lane, York YO10 5GB, UK. E-mail: sandra.pauletto@york.ac.uk

(Received 27 February 2013; accepted 24 May 2013)

ISSN 1401-5439 print/ISSN 1651-2022 online
DOI: 10.3109/14015439.2013.810303

as well as the length of pauses and variations in speech flow.

A speech synthesizer is capable of simulating these acoustic contributors with varying degrees of fidelity (10). Emphasis is on the word *simulate*. Popular language uses the word 'emotion' in technology as though it were identical to human emotion. But, of course, a computer has neither a biological body nor consciousness and is therefore not actually feeling any subjective experience.

In this paper, we use the three terms *affect*, *feeling*, and *emotion*, as defined by Shouse (11), to distinguish between preconscious physiological changes such as galvanic skin response, heart/breathing rate, and/or hormonal changes and their biological intensity (affect); the subjective perception and conscious integration of embodied states (feeling); and the projection/display of feelings (emotion). This third display component is what is in play with text-to-speech (TTS) synthesis. That is, an artificial voice *expresses emotion* in the sense that the measurable acoustical effects of emotional speech, as produced by a human vocal tract, are replicated in the acoustical output of the artificial voice. Most speech synthesizers support, for example, specific control parameters for the known acoustic contributors to emotion. When they are modulated, the synthesizer is *simulating human emotion* in the real sense that its goal is to communicate with a human listener and that the information communicated is presumed to be emotional information, separate from the syntactic and semantic information contained in the phonology of the speech. This is the explicit assumption of the standard content paradigm. Different technologies exhibit varying degrees of fidelity, measured through reports by human listeners and their alignment with stated intentions by the supervisors of the technology.

Several interrelated terms associated with the study of emotion are shown in Figure 1 as a tag cloud. These terms vary greatly between disciplines, including physiology and psychology as well as general art and science. A larger font size indicates greater relative importance to the VEEG project.

For this project we are dealing with artificial voice capabilities in a general way and so have adopted the simplification of merging the disparate terms into two very broad categories: 1) those phenomena that can be externally observed and measured, described primarily as events in the body (Figure 1, black font); and 2) those phenomena that are internal/subjective, exist in terms of perception and consciousness, are more difficult to measure physically, and are described primarily as events in the mind (light grey font). 'Feeling' is the most significant word in this category.

Figure 1. Tag cloud of related terms in alphabetical order.

We use the term feeling to mean subjective events in the mind. That is, the complex and multilevel phenomena that underlie human perception and interpretation of the external world, along with the emotions that they trigger, are both cause and effect, and involve extremely complex feedback loops: sensation and perception modulate affect and its subsequently displayed embodied emotion, which in turn changes the perception of the body-state and its attendant feelings. Machines display emotions but do not feel them. Only biological organisms can feel emotions.

The dark grey font of Figure 1 represents the movement of emotional information between machine and human. Here we distinguish between expression, communication, and conversation.

For our purposes, expression is a one-way movement of emotional information (i.e. broadcast), while communication is the bidirectional exchange of such information. In the case of the former, the absence of a feedback loop makes it intrinsically less complex and therefore less interesting, but also makes it more tractable.

In addition to acoustic changes between different emotions, we need to consider the variations in the expression of the same emotion, visible as levels of arousal. Bänzinger and Scherer (12) make the distinction between two main levels of emotional arousal. High arousal, for example hot anger and elated joy, are associated with a louder voice, fast speech rate, and higher pitch. This contrasts with low arousal emotions, for example cold anger and calm joy.

Although the human emotional space is vast and includes large variations between people, it is possible to distinguish a limited number of emotions that have common characteristics in all human beings. These are called *basic emotions* and are the starting-point of most attempts to add emotions to a synthesized speech signal. Examples are anger, sadness, and happiness. Basic emotions are often viewed as the root of more complex emotions, and it has been suggested (13, in 14) that the prosodic content—which according to Murray and Arnott 'consists of three main elements: amplitude structure (including stress and prominence), temporal structure (pause, rhythm, and segment duration) and pitch structure (accent and intonation)' (14, p. 109)—creates the primary

distinction between basic emotions, while voice quality, or the energy distribution in the speech signal spectrum, allows us to distinguish between complex emotions.

In addition, the distinction mentioned between high and low arousal states is consistent with Ekman's view (15) that each emotion belongs to 'a family of related states' (15, p. 55), which, depending on what is experienced, produce varied emotional results.

In this study, we focus on emotions, not feelings, and attempt to imply emotion through the manipulation of published and standard TTS technology parameters. Our observations attempt to determine whether the emotions inferred by listeners correlate with those implied by us.

Material and methods

For the initial CreST project leading up to the roadshow, the VEEG goals were modest:

- Perform some initial experimentation to understand technology materials;
- Build a simple device that allows hands-on experimentation;
- Develop hypotheses in the form of questions about artificial emotion;
- Use the device to refine the hypotheses and pursue more subtle areas of interest; and,
- Demonstrate the device in public for feedback and observation.

With these goals in mind, we built a chatbot, an artificial conversational software device, that could allow real-time human–computer interaction (HCI) in the form of conversations that could stimulate emotions in the human and give us opportunities for experimenting with emotion design in synthesized speech.

Because of its specific design as a 'characterful' speech synthesizer (10), we chose the CereVoice® TTS software produced by CereProc Ltd (16), an Edinburgh technology vendor. Initially, we used Balabolka (17), a TTS playback engine downloaded free from the internet, to render a variety of SAPI-compliant voices, saving them as .wav files. The acronym SAPI stands for 'Speech API', a standard software interface supported by Microsoft and others. For initial experimentation, examples of text were developed reflecting hypothetical emotional categories, and then TTS samples were recorded using various emotion tags. Several early (SAPI version 4) TTS voices were compared with newer (SAPI version 5) voices. Most of the final examples used Heather, a Scottish CereVoice character. These examples are posted on the internet along with a discussion (18).

Tests were then carried out to find an appropriate method for navigating the emotion space to simulate the emotional changes of our 'digital entity'. One common two-dimensional mapping, shown in Figure 2, plots emotional valence along the *x*-axis from negative to positive. The intensity of the emotion is described in terms of arousal along the *y*-axis.

CereProc allows selection from among three sets of sampled units. The primary set, typical of unit selection technologies, represents a *neutral* emotional state. Two additional sets then provide the base units for emotional states, *stressed* and *calm*. From the *stressed* units two emotion tags are derived which are *crossed* and *sad*, while from the *calm* units the emotion tags *calm* and *happy* are created. These emotion tags are characterized by imposing specific settings of acoustic parameters such as pitch, amplitude, and speech rate onto the base unit set.

The pilot test consisted in applying these emotional tags at appropriate insertion points in an emotional text. This approach, however, resulted in a disjointed speech, with audible discontinuities in acoustic characteristics which were not appropriate for our speaking 'digital entity' (chatbot).

To solve this problem, we have created a path in the emotion space that takes us in small steps from an extremely negative emotion (very cross) to an extremely positive one (very happy). This path evolves between negative and positive emotions passing through the emotional states *cross*, *sad*, *neutral*, *calm*, and *happy* and their relative acoustic settings. We have created a 13-step scale in which 0 represents the *neutral* state, + 6 the *very happy* state, and –6 the *very cross* state. The acoustic settings of the intermediate steps were found initially by linearly interpolating between the given acoustic settings of the emotion tags, and then optimized by listening to the results and adjusting the acoustic parameters. The scale was tested using short, emotionally meaningful, sentences shown in Table I.

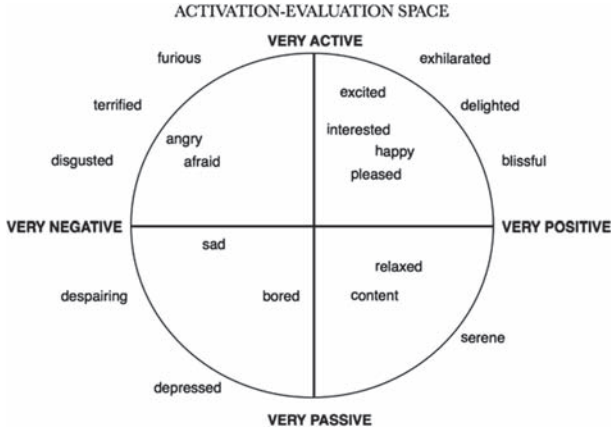

Figure 2. Emotion plotted as valence and arousal.

Table I. Emotional scale with example sentences.

Emotion	Sentence
+6 Very happy	What a fantastic idea! Great! I am so excited we found this solution.
+5 Happy	Yes, I completely understand. That sounds great to me. Yes, that's brilliant I know. I completely know what you mean, yes.
+4	That sounds good. A very good idea. Well done!
+3	Good idea, yes, sounds good!
+2 Calm	Yes, I understand, that sounds good to me. Yes, thanks, I know. I really know what you mean.
+1	Yes, good, a good idea.
0 Neutral	I am not sure I understand completely.
−1	No, I don't think I understand. It seems confusing to me.
−2 Sad	No, I am not sure I understand, that sounds confusing to me, no, I don't know, I don't know what you mean.
−3	No, I really don't think I can understand what you are proposing. It's quite confusing to me.
−4	I don't quite understand what you are saying. It's quite confusing and I find it too unclear.
−5 Cross	I don't understand what you mean. That is really confusing me. No, I don't know what you are talking about, I don't have any idea what you mean.
−6 Very cross	That is just not good enough! It's confusing and unintelligible. You should have thought about it properly!

In addition to this 13-step scale, two extreme states were created at each end of the scale (+7, –7), characterized by non-human acoustic parameters. When the chatbot adopted these extreme states, we whimsically referred to it as 'mad'. Human–chatbot conversations that remained in this state were classified as 'rubbish' in the software.

This, now, 15-step scale allows us to evolve a text from one emotion to another in a relatively smooth way. The main limitation of this method is that all the steps need to be applied when moving from one emotion to another. This means that the text needs to be long enough to allow for a smooth transition. To accomplish this, sequences of verbose text were composed that contained ambiguous emotional cues (19). The text constituted a baseline that eventually evolved into some of our extended chatbot monologues. This requirement for long text became particularly limiting, however, when needing to transition rapidly between two emotions at the ends of the scale. So additional design work, focused on splitting monologues into dialogues, allowed us to spread the emotional transitions across multiple turns between human user and chatbot. Future work on different ways to navigate the emotional space, depending on how quick the emotional transition needs to be, could give more insight into this issue. The essential observation is that human emotions do not spring suddenly into existence on this or that word; they evolve continuously over time. More research is needed to understand this transformation and how to represent it using moment-in-time emotion tags that are interspersed directly into the text.

In addition to the emotion tags, we ran tests on so-called speech gestures, i.e. non-speech vocal sounds such as 'oh' positive or negative, or 'sigh' sad or happy. Results showed that they were very helpful in reinforcing the emotional content of a speech stream. However, repetition, in a short space of time, of the acoustically identical speech gesture diminishes the emotional impact of the speech. We concluded therefore that emotional gestures need to be varied acoustically, especially when the automated addition of emotional gestures is desirable.

The chatbot as a conversational kiosk

The VEEG artefact is conceived as an installation, such as in a museum or public space. The installation consists of one or more kiosks, a sort of physical embodiment for the artificial entity. Each kiosk houses speech input and output technologies, containing, at a minimum, a microphone delivering human speech to a speech recognizer, and a TTS synthesizer connected to a loudspeaker. Together these technologies constitute the ear and the voice of the kiosk.

A chatbot is an artificial conversational software device. They are commonly used today for customer support applications (using typed text), and much research has been devoted to chatbot design in the context of Turing tests and The Loebner Prize (20). The VEEG implementation can be thought of as a kiosk that is running chatbot software, but for simplicity and readability, we herein use the terms kiosk and chatbot synonymously.

The minimum implementation, built for the CreST *articulate* roadshow in late 2012 and early 2013, is a single kiosk demonstrating chatbot conversations. Subsequent work will build on this minimum implementation as VEEG membership and funding allow.

The highest-level design for the VEEG chatbot is shown in Figure 3 as a state-transition diagram.

After initialization at state A.1, the chatbot enters a quiescent state (A.2) and waits for input. A conversation is triggered at 5 through the press of

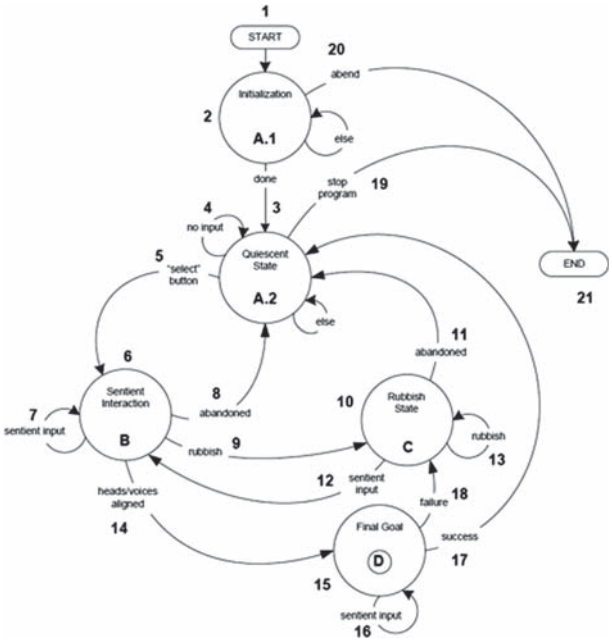

Figure 3. Chatbot state-transition diagram.

a button. The chatbot takes the first turn at 6, speaking some prompting phrase through the loudspeaker. The automatic speech recognition (ASR) is then activated, and the user may speak. As long as user speech is 'sentient' (represented in the state table as calling for a specific response), the system remains in state B as shown at 7. Sentient speech output can present a low-arousal to medium-arousal state on the emotional scale of Table I.

When the user deviates from a pre-defined dialogue at 9, the chatbot enters a new state at C called 'rubbish'. In this state, the spoken text becomes less relevant to any sentient conversational goal, and the synthesizer begins exhibiting more extreme emotional attributes. Remaining in state C for any length of time gives the appearance that the machine has gone mad (in terms of both semantics and emotion). User input may or may not exhibit any kind of emotional reaction to this condition, and the ASR has no way of knowing what the user is saying or meaning. At this point, all 'communication' is only in the user's imagination (an emergent property of the coincidences of the conversation).

The chatbot software was written in C++ using Microsoft Visual Studio on a Windows 7 PC (21). We chose Lumenvox speech recognition software (22). The following components were installed on each of three laptop computers to allow concurrent testing during development:

- One local license of the CereVoice TTS Engine;
- One local license of the Lumenvox ASR engine;
- Several CereVoice voices with emotional capabilities;
- One copy of Balabolka; and,
- The chatbot software and related text files.

Results

There are four results to report: 1) observations of the kiosk/chatbot in use; 2) questions that emerged during the investigation; 3) sketches of user interface designs for markup tools; and 4) reactions by the public to the pilot (beta test) and the formal roadshow.

Result #1: The kiosk/chatbot

In a public performance mode, the design allows actors to approach the chatbot and engage in a conversation with it. The conversation is scripted but allows for deliberate (improvisational) or random deviations.

In a public display mode, spectators from the general public approach the chatbot and listen to it speaking. They may or may not engage in conversation. If they do, the conversations will be completely unscripted and depend on the internal dialogue model of the kiosk, as designed by VEEG. This makes the installation essentially an interactive art exhibit.

The physical appearance of the kiosk is of secondary importance for the initial roadshow. Figure 4

Figure 4. Kiosk/chatbot setup for beta test.

shows the physical appearance of the VEEG stand for the beta test pilot in September, 2012. During setup (a), four of us (from left to right, B.B., L.B., D.P.M., and M.A.) collected an assortment of found objects (junk) into our assigned footprint in the Ron Cooke Hub at York University. Software tweaking and sound-level adjustments (b, from left S.P., B.B., and D.P.M.) and arguing (c, from left, B.B., S.P., L.B., and D.P.M.) then ensued. Figure 5 shows the resulting kiosk in use (from left to right, a spectator/user, K.J., and S.P.). The microphone is hidden inside the motorcycle helmet, and the loudspeaker behind the white stand presents TTS speech. Note that the lamp is lit, indicating that the chatbot is speaking.

The physical configuration for the kiosk varied throughout the roadshow according to variations in the venue, including the acoustical environment, spectator traffic patterns, and practical considerations. Shown in Figure 6, the motorcycle helmet and lamp (a) served as an abstracted 'face' in some cases, while a laptop running *Soundstream* (b), a sound-responsive screen saver (23), took on that role in others. The simplest exhibit consisted simply of a telephone handset (c).

Result #2: Working questions

The following questions emerged during the course of the study:

- What is the best voice for a given application?
- Whether digitally recorded or synthesized, how should an artificial voice sound to its user? Can it express emotion? Should it?
- Designers of voice applications often speak of personality. What does that mean in reference to an artificial voice? What role does the personality play in the user's interaction with the voice?
- Can a computer ever be said to have feelings? Is it desirable? That is, is there any applied reason for a computer to not only display emotions but to feel them subjectively?
- When a voice is 'acting', that is, producing effects aimed at convincing a listener that feelings, emotions, intentions, or personality traits are present when in fact they are not, do listeners respond as though the machine were a real human actor?
- Can the performance and emotional abilities of such a digital voice be used for other artistic endeavours? Can this voice be of use in musical composition and singing?
- Can acting by an artificial voice be beneficial in an HCI context, an artistic context, or a marketing context?
- When an artificial voice is serving as the primary communication medium for a human user, for example people who have lost their voices or do not have adequate motor control for intelligible vocalization, does the voice include identity cues? That is, can the voice be thought of as representing that individual? Is that individual in fact 'speaking through' the artificial medium? If so, how does that change the answers to these questions about emotion?
- What sort of high-level tools do we wish for? Are these tools different between disciplines?

Figure 5. Chatbot/kiosk in use.

Result #3: User interface sketches

The most difficult problem for the entire VEEG team was the absence of intuitive tools that support high-level text markup. Most of our markup used Speech Synthesis Markup Language (SSML), a published, open standard (24). But entering specific SSML tags directly into the running text and then generating .wav files for comparative listening proved tedious and time-consuming. We found ourselves wishing for a user interface layer above the tags, as shown in Figure 7.

The interface would serve as an abstraction that allows designers to think intuitively about the problem. There are several ways such a design might be approached; following are a few. It is our hope that technology engineers will be interested in taking these simple sketches and turning them into a full-fledged detailed specification for some suite of future TTS management software.

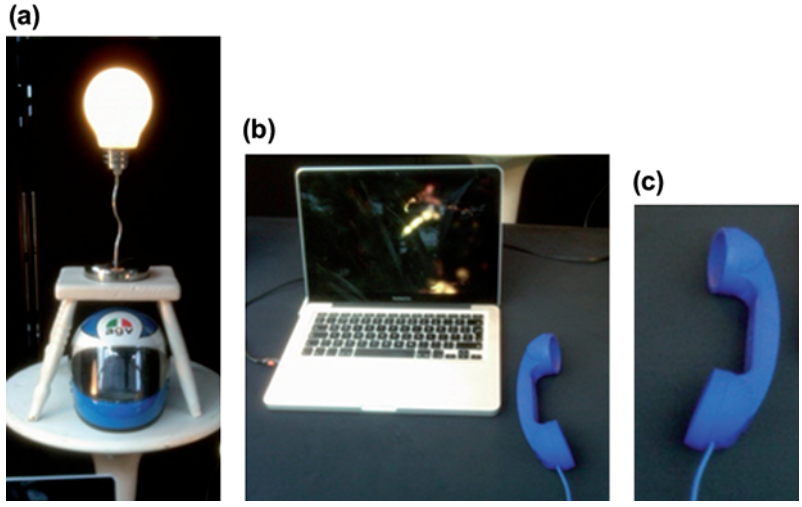

Figure 6. Design variations for roadshow.

Interactive text markup. One simple approach to specifying key prosodic elements is the direct manipulation of text shown in Figure 8.

The text editor should support reading, displaying, selecting, modifying, and saving raw and marked-up text residing in text files with no special format. Such a tool would be simple to build and very useful for the designer. In the example, font size represents 'stress', a combination of pitch and loudness attributes. The interface would map the font-size parameter directly onto SSML tags, either through computation or by table lookup. This solution would adequately serve most needs of playwrights, commercial voice announcement design, and some HCI requirements.

Somewhat more complex graphical manipulators, shown in Figures 9 and 10, allow separation of the loudness and pitch parameters. Pitch is specified by the relative placement of text vertically between minimum and maximum values.

In Figure 9, the user may click on the centre-line to edit the centre pitch for f0, shown here as 110 Hz. The top and bottom lines represent degree of deviation from centre in Hz, influencing users' perception of arousal. Loudness is specified by font size. Simple and easily available symbols such as the underlines (two are shown) may be used to represent attributes such as unit sets or amplitude envelopes. Rate of speech changes might be specified with slider widgets.

In Figure 10, more elaborate visual objects point towards prosodic patterns of various kinds. Controls, similar to those in Figure 9, are not shown but can be easily visualized for independent manipulation of voice tags with effective corresponding visual correlations that are at once intuitive and fast to use. The sketches shown in Figures 9 and 10 have applicability to augmentative and alternative communication (AAC), HCI, and voice announcement applications.

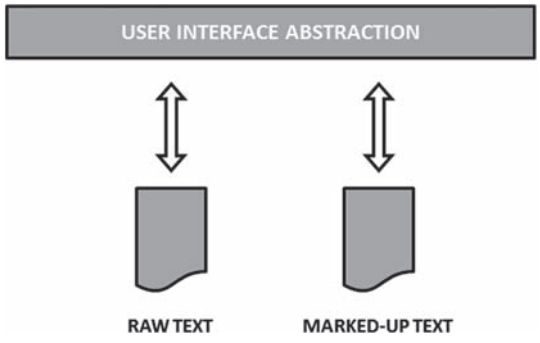

Figure 7. User interface for specifying emotions.

Parametric control. Graphical controls for manipulating parameters might include such standard devices as rubber-band lines and spline functions. A simple example is shown in Figure 11. The user drags the attachment-point handles up or down to vary the value of a given parameter, for example combined pitch and loudness (an approximation of stress).

Real-time control. Control in real-time may take advantage of keyboard, joystick, pressure-pad, or similar off-the-shelf control devices. A MIDI keyboard, for example, is an effective interface for 'playing' the TTS, manipulating pitch, loudness, and other parameters dynamically and even improvisationally. It should be noted that lag times inherent in such control make the term 'real-time' problematic. SSML tags must be dropped into the text stream before the target phrase begins, and influencing speech that is currently being heard by the performer is difficult to implement. In particular, certain TTS

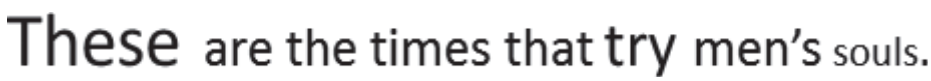

Figure 8. A simple text editor for stress markup.

technologies are more amenable to real-time manipulation than others. More research is needed to determine feasibility of real-time control.

Musical composition. The most complex but desirable user interface for musical composition requires a deeper integration with contemporary musical notation software. Figure 12, for example, shows a snippet from a musical score of the proposed part for a singing TTS voice.

In the example, a composer would write for the TTS in the same way as for any speaking or singing voice. Round note-heads (a) indicate specific pitches and X note-heads (b) indicate approximate pitches. In some imagined integration of the speech synthesizer with the composition software, the text would be converted to audio by the TTS synthesizer in the same way that the other musical notation is converted to MIDI.

Presumably, the conversion would pass through a MIDI representation stage, or conversely would be rendered directly into SSML tags that are inserted into the text to be interpreted by the TTS front end. Approximate pitches might convert into prosodic contour or approximate (relative) pitch tags; specific pitches would be represented in precise equal-tempered designations in Hz.

Result #4: Reactions by the public

The conversational kiosk was presented to the public during the fall and winter of 2012/2013 in the North of England.

- 19 September: York University (beta dress rehearsal)
- 3 December: York, City Screen
- 4 December: Sheffield Winter Gardens
- 5 December: Hull Truck Theatre
- 23–26 January: Scarborough Woodend Creative Centre

As expected, some spectators were tentative, saying things like, 'I don't know what to do; Am I supposed to talk to it? How long do I listen? What am I listening for?' These people tended to listen for a short time without interacting. This seemed to be the case even when the chatbot was asking for input ('Are you there? Do you agree?').

Others jumped right in, listening and then speaking at appropriate junctures. These people tended to chat for a long time. Unsurprisingly, the longer they spent with the chatbot the more they adapted to the turn-taking, and the more accurate the speech recognizer became. These users seemed to smile and laugh the most often at the content, especially the jokes.

The observation that interaction increases social engagement, which in turn increases both perceived and real technology performance, is well-known and commonly observed in other artificial speech dialogues. The phenomenon derives from the personality of the participant (e.g. willingness to suspend disbelief) as well as the set and setting of the conversational environment.

A second observation, equally well-known, is that users form a theory of the inner workings of the machine (Theory of Mind) and then adjust their behaviour and form opinions based on that theory. The following sample quotes point towards this user tendency:

> It's hearing keywords, and then going on. Like it grabs an idea and goes with it. (York, 3 December 2012)
> I tried to mislead it but it wouldn't allow me to. As long as I follow its lead, it responds to me. But If I go off-script, it pulls me back. (York, 3 December 2012)
> He just sounds like a politician—avoiding a question. (Hull, 5 December 2012) [This one was in response to the chatbot saying, 'that's a very good question; it reminds me of a story.']
> He's ranting now. [This comment was spoken by several people after the chatbot transitioned to one of its 'mad' states.]

Although the roadshow was a public demonstration and not a scientific research project, we can take

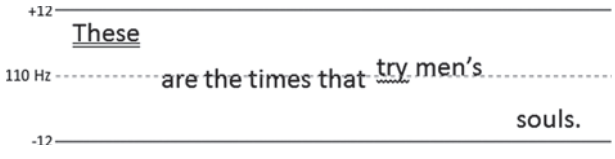

Figure 9. More elaborate prosodic specifier.

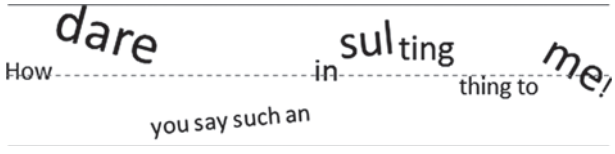

Figure 10. Graphical text manipulation.

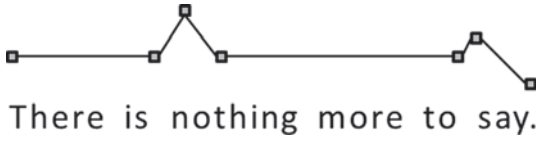

Figure 11. Parametric control.

from these observations a conclusion that human users do tend to respond emotionally to spoken displays, and in fact tend to inject emotional implications into even random spoken dialogues. The challenge therefore becomes how to manage machine emotion and how to exert emotional control over such dialogues.

Discussion

The goals of this investigation were interdisciplinary, with interest in musical composition, augmentative and alternative communication (AAC), commercial voice announcement applications, human–computer interaction (HCI), and artificial intelligence (AI). In this discussion we touch on each of these disciplines.

Composers and sound artists are faced with practical and aesthetic dilemmas, all of which share, with speech technology, a specific end-point: a human listener hears an audio signal. As in any problem-solving process, the first and defining step is to cast the question into a form that points towards appropriate means for answering it. The audio world boasts an abundance of tools, constantly changing and expanding, that can be thought of as various 'languages' for depicting, generating, and manipulating sound. These languages range from conventional music notation, through MIDI mapping and waveshape editing, to digital recording environments, among others. These languages represent different ways of asking the questions that compositional decisions answer.

Languages which serve to depict sound are in a constant state of flux, morphing to accommodate the similarly morphing questions posed by the artists using them. The notational system used by Guillaume Dufay was a language appropriate to the needs of a fifteenth-century composer, while the notational system used by John Cage was a language appropriate to the needs of a twentieth-century composer. The C++ programming language, the Apple Logic, or Avid Pro Tools recording environments, and contemporary music notation are all languages that are appropriate to the needs of twenty-first-century composers. Digital plugins, a kind of subset of languages, designed to perform highly specific audio tasks, multiply on a daily basis.

However, none of these languages are exclusive. The appropriateness and usefulness of any language is determined by the nature of the questions being asked. Composers, like the practitioners of other VEEG disciplines, work simultaneously in several languages. We often observe ourselves struggling mightily, trying to force one language to address a question, before realizing (sound of hand slapping forehead) that the problem is actually solved rather easily by using a different tool or language. 'If all you have is a hammer, everything looks like a nail.' If one is working only in conventional music notation, every question appears to be about notes on a staff; if one is working only in C++, every question appears to be a programming question.

Designers and users of AAC devices have certain overarching requirements, e.g. cost, ease-of-use in the face of missing or impaired modalities, real-time control, ease-of-customization/personalization, and ruggedness. Aesthetic or amusing expressivity is of little interest, but miscommunication caused by attenuated or distorted emotional expression is of great concern. The AAC community also stands to benefit from answers to questions about TTS and self-identity.

In conclusion, we must ask ourselves what our questions really are. If we want to find spoken examples of authentic human emotion, then the answer is to find human beings feeling emotion and then listen to them speak. This non-technology solution is fully mature today for playwrights, poets, composers of song, and everyday life.

If the question is how to make pre-generated speech sound the way we want it to sound, then the answer may very well be to record speech (either with human actors or TTS) and then post-process the audio signal with audio tools. This solution is perfectly acceptable for in-studio composers, commercial voice announcement messages, certain HCI

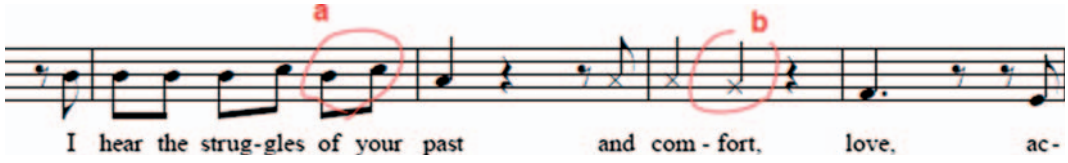

Figure 12. Composing vocal music.

applications (including interactive voice response), and even certain limited AAC applications.

If the question is how to modulate the prosodic-generation algorithms inside a TTS synthesizer, modifying known text to generate emotional speech automatically at some unknown time in the future, then a high-level interface that inserts tags into the text as described herein may be a good answer. This solution serves the same audience as the preceding one.

If the question is how to improve prosodic generation such that unknown text will exhibit improved emotional fidelity without human intervention, then the tools described here will not help, and are, in fact, answering the wrong question. Better questions, deriving perhaps from the AI community, might uncover better answers, hopefully serving the AI community and their users.

Whatever language we might prefer or are most comfortable with, one must always remain open to the possibility that there may be another, more efficient language for the task at hand, whether that language has its roots in the fifteenth or the twenty-first century. But if we fail to ask the right questions, we may very well end up with human users and machines, as shown in Figure 13, staring in dumb incomprehension at each other across an unbridgeable void.

Note that, in most of these application areas, the technology itself is a proxy for the human supervisors and cannot be said to be experiencing any emotion (since it has no body) nor any feeling (since it has no mind). For this reason, the emotion being expressed is encoded by the human supervisor for her own reasons. This is an important point. As an AAC device, an artificial voice expresses on behalf of its user. As an acting or singing artist, the voice expresses on behalf of the poet, playwright, or composer. A public announcement or alert expresses the intent of the authorities. Regardless of application, the technology is serving as a proxy for human emotions which it does not and cannot feel.

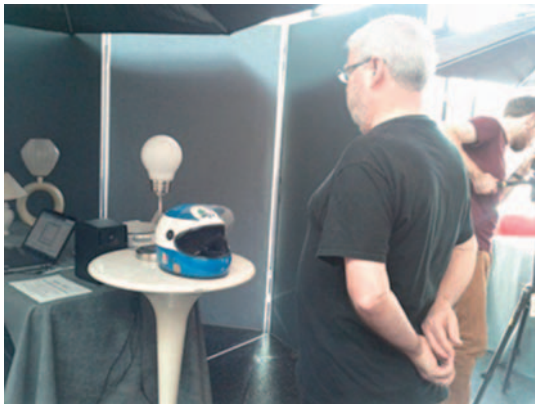

Figure 13. A conversational stand-off.

This gets to the heart of the CreST Network mission. When humans complain that a synthesizer sounds 'robotic' or 'alien', the problem is not that there is no emotion contained in the signal. The problem is that the signal is expressing the wrong emotion, leading to confusion and miscommunication. Our VEEG goal is to discover whether it is possible, and if so how, to make an artificial voice accurately and faithfully express its human supervisor's emotional intention.

Acknowledgements

We would like to thank CereProc and Lumenvox, who donated their technologies for this effort.

Declaration of interest: The authors report no conflict of interest.

We also would like to thank all colleagues from the EPSRC CreST Network for funding and supporting this work.

References

1. Articulate: The Art and Science of Synthetic Speech, http://crestnetwork.org.uk/ [Last Accessed: 28 June 2013]
2. Scherer KR. Expression of emotion in voice and music. J Voice. 1995;9(3):235–48.
3. Davitz JR. Personality, perceptual, and cognitive correlates of emotional sensitivity. In: Davitz JR, editor. The communication of emotional meaning. New York: McGraw-Hill; 1964. p. 57–68.
4. Scherer K. Personality markers in speech. In: Scherer K, Giles H, editors. Social markers in speech. London: Cambridge University Press; 1979. p. 147–210.
5. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J Acoust Soc Am. 1993;93(2):1097–108.
6. Schröder M. Emotional speech synthesis: a review. In: INTERSPEECH, Paul Dalsgaard, Proceedings: Eurospeech 2001, Scandinavia, 7th European Conference on Speech Communication and Technology, September 3–7, 2001, Aalborg Congress and Culture Centre, Aalborg, Denmark, Volume 3.
7. Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull. 2003;129(5):770–814.
8. Pauletto S, Bowles T. Designing the emotional content of a robotic speech signal. In: Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound, Piteå, Sweden – September 15–17, 2010, Available at: http://dl.acm.org, (Last Accessed 28 June 2013).
9. Bowles T, Pauletto S. Emotions in the voice: humanising a robotic voice. Proceedings of the 7th Sound and Music Computing Conference; Barcelona, Spain, 21–24 July; 2010 Available at: http://smcnetwork.org/files/proceedings/2010/30.pdf [Last Accessed 28 June 2013].

10. Aylett MP, Pidcock CJ. The CereVoice characterful speech synthesiser SDK. Newcastle: AISB; 2007. pp. 174–8.
11. Shouse E. M/C Journal: A Journal of Media and Culture, Volume 8, Issue 6, 2005. Available at: http://journal.media-culture.org.au/0512/03-shouse.php [Last Accessed 28 June 2013].
12. Bänzinger T, Scherer K. The role of intonation in emotional expressions. Speech Communication. 2005;46:252–67.
13. Stockholm. Available at: http://www.speech.kth.se/music/publications/kma/papers/kma33-ocr.pdf [Last Accessed 28 June 2013].
14. Murray IR, Arnott JL. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Computer Speech and Language. 2008;22:107–29.
15. Ekman P. 1999. Basic emotions. In: Dalgleish T, Power M, editors. Handbook of cognition and emotion. Chichester: John Wiley & Sons. p. 45–60.
16. CereVoice Engine Text-to-Speech SDK, http://www.cereproc.com/en/products/sdk, [Last Accessed 28 June 2013]
17. Balabolka Software Download, http://balabolka.en.softonic.com/, [Last Accessed 28 June 2013].
18. Available at: http://www.eiginc.com/ResearchFiles/SpeechMusicContinuum/index.html.
19. Jones K. Unpublished monologues. 2011. Available at: CreST.
20. Christian B. The most human human. New York: Anchor Books; 2012.
21. Windows 7 Download, http://windows.microsoft.com/en-GB/windows7/products/home, [Last Accessed 28 June 2013].
22. LumenVox Software, http://www.lumenvox.com/, [Last Accessed 28 June 2013].
23. Soundstream: sound-responsive screensaver, http://pcheese.net/software/soundstream/, [Last Accessed 28 June 2013].
24. Speech Synthesis Markup Language (SSML) Version 1.0, http://www.w3.org/TR/speech-synthesis/, [Last Accessed 28 June 2013].