SpeechPlay: Composing and Sharing Expressive Speech Through Visually Augmented Text

Kian Peen Yeo[1], Suranga Nanayakkara[2]
Augmented Senses Group
Singapore University of Technology and Design
20 Dover Drive, Singapore 138682
yeokianpeen@sutd.edu.sg[1], suranga@sutd.edu.sg[2]

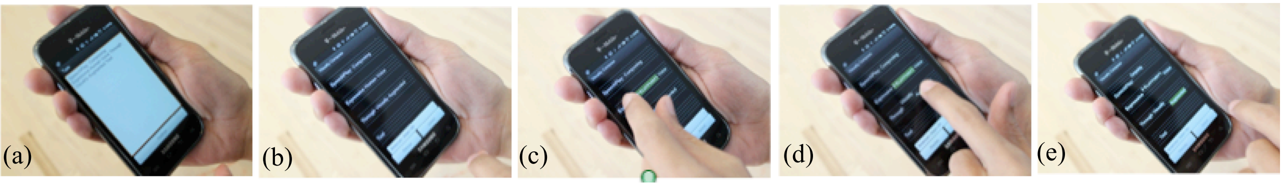


Figure1: SpeechPlay – (a) Type (b) Compose (c) Pinch scale (d) Drag (e) Hear.

ABSTRACT

SpeechPlay allows users to create and share expressive synthetic voices in a fun and interactive manner. It promotes a new level of self-expression and public communication by adding expressiveness to a plain text. Control of prosody information in synthesized speech output is based on the visual appearance of the text, which can be manipulated with touch gestures. Users could create/modify contents using their mobile phone (SpeechPlay Mobile application) and publish/share their work on a large screen (SpeechPlay Surface). Initial user reactions suggest that the correlation between the visual appearance of a text phrase and the resulting audio was intuitive. While it is possible to make the speech output more expressive, users could easily distort the naturalness of the voice in a fun manner. This could also be a useful tool for music composers and for training new musicians.

Author Keywords
Speech Synthesis, Mobile Devices, Interaction Design, Visual Communication.

ACM Classification Keywords
H5.2. User Interfaces

INTRODUCTION

Speech synthesis technology is found in various applications, more often as an assistive tool for people with a wide range of disabilities [8]. In more recent years, with improving quality in the naturalness of synthetic speech, the technology has found use in other applications including telecommunications, multimedia and robotics [4, 12]. Most research on speech synthesis works towards the ultimate goal of an artificial system that generates truly natural and expressive vocal output [19]. Typical 'user-friendly' text-to-speech (TTS) systems allow users to change synthesizer settings through a Graphical User Interface (GUI) that has drop-down menus, sliders etc. However, with these GUIs, it is difficult for users to control the various settings and have an idea of the resulting synthesized voice. Text by itself does not convey auditory expressiveness when synthesized. However, symbols in the form of musical notations are able to provide a visual representation of aurally perceived tones. In fact, a number of phenomena such as the ventriloquism effect [11], 'McGurk effect' [16] and synesthesia [6] demonstrate how auditory and visual information can mutually reinforce or modify sensory perception. It has also been shown that the human peripheral cortex helps to bind the major aspects of audio-visual features to provide meaningful multi-modal representations [21]. Thus, we believe that visually augmented text could become a powerful tool for encoding auditory expressiveness. In addition, touch-enabled public displays have become mainstream in recent years and smart mobile phones have become a common accessory. In this work, we focus on enhancing social communication by combining these two popular media (mobile phones and large displays).

We do not plan to go into the details of the underlying technology behind TTS and emotional speech synthesis. Instead, we intend to focus on highlighting the innovative ways of how this technology has been applied to interactive applications involving the use of visuals and/or gestures to manipulate synthetic speech. Blankinship and Beckwith [3] developed a standalone computer application called *Poet Shop* which allows a user to modify the volume/pitch and the spoken rate of synthesized speech through a contour plot and manipulation of text typography. In an earlier work on a speech to text application by Rosenberger and MacNeil [18], speech was represented by animated fonts that reflect the emotional aspect of its prosodic contents. Both

OZCHI'13, November 25–29, 2013, Adelaide, SA, Australia.
Copyright 2013 ACM 978-1-4503-2525-7/13/11…$15.00.
http://dx.doi.org/10.1145/2541016.2541061

these work suggest the important link between text typography and prosodic contents in expressive speech. Alessandro *et al.* [7], in their interactive voice synthesis projects, DiVA and HandSketch, proposed a new method of voice synthesis through the use of gestures without textual input. While their proposed systems achieve a new form of interactivity, they were complicated to use and required a considerable level of technical expertise to operate [9].

There has been a number of works exploring the use of mobile phones and public displays as a medium of self-expression and public communication [1, 2, 5, 15]. While most of these systems generate visual expressions based on texting (SMS), images or computer graphics, they however do not experiment with auditory feedback. Our fingers have an incredibly rich and expressive repertoire [14] which allows us to manipulate objects without even thinking about it. Thus, common gestures such as tap, drag and pinch have become a natural way to interact with touch-enabled user interfaces. In SpeechPlay, we use these common touch-based gestures to manipulate the visual appearance of a text input that will encode expressiveness into a synthesized voice, which can be then shared with other users.

SPEECHPLAY SYSTEM

The idea behind SpeechPlay is to provide a 'composition playground' enabling users to synthesize expressive human voice. We want to provide a private space for the users to experiment with composing and a public space to showcase their work and collaborate with others [10, 19]. Thus, we developed a system consisting of a mobile phone application called SpeechPlay Mobile and a large interactive wall called SpeechPlay Surface (Figure 2).

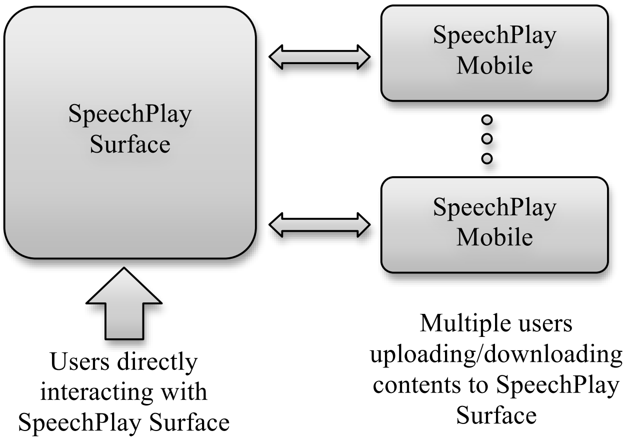

SpeechPlay Surface
SpeechPlay Mobile
SpeechPlay Mobile
Users directly interacting with SpeechPlay Surface
Multiple users uploading/downloading contents to SpeechPlay Surface

Figure 2. Overview of the SpeechPlay System

Users with SpeechPlay Mobile installed on their mobile device first create or modify an existing message. Through different touch gestures, users can then augment the text to hear a resulting synthesized speech output of the message. Once satisfied, SpeechPlay Mobile users within the vicinity of the SpeechPlay Surface can upload their creative compositions to the display. SpeechPlay Surface provides an interactive interface to display some of the recently uploaded or viewed text compositions. The synthesized speech can be played back to anyone who is browsing through these compositions on the touch-enabled display. Users can select a specific composition to hear it or copy it from the SpeechPlay Surface and paste on to his/her SpeechPlay Mobile application and make modifications to the original composition. These interactions encourage user participation by collaborative play [17] and *active spectatorship* [12]. The current system has been implemented only on English language.

Text Typography and Mappings

The SpeechPlay system provides an intuitive mapping between text typography and prosody information in synthesized speech. A word spoken at its default rate will be displayed at the default size and spacing. Stretching or compressing a written word will change its horizontal width, thus increasing or decreasing its spoken duration (Figure 3a). The motivation for visual representation of pitch (frequency) comes from the concept of musical notations. The musical staff line and space notation together with the position of words (representing the musical note symbol) are used for visual representation of pitch. The default position is at the middle line, and words can either be on the line or space, giving a total of 11 possible pitch selections. Positions above and below the middle line represent an increase and decrease in pitch respectively (Figure 3b). From a typographical point of view, in text representation, bold or larger fonts are typically used to emphasize words such that they stand out from the rest of the text. Following this general idea, we use a mapping of bold formatting on a written word to represent an increase in volume for its synthesized output (Figure 3c). Finally, the duration of pauses between spoken words is mapped to the horizontal displacement between written words (Figure 3d).

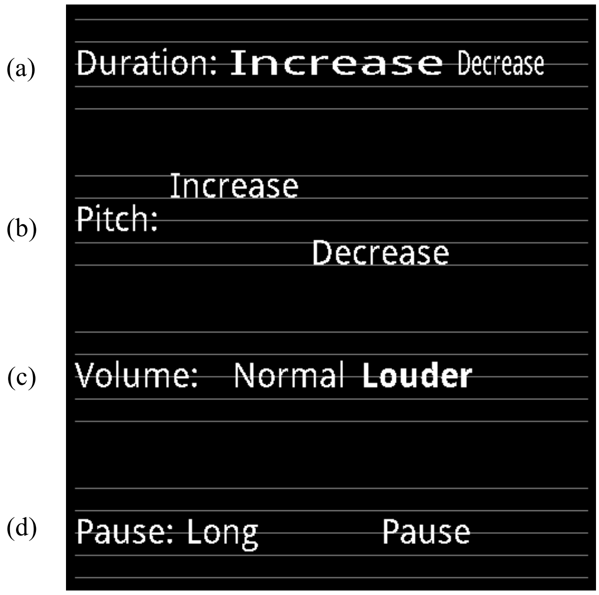
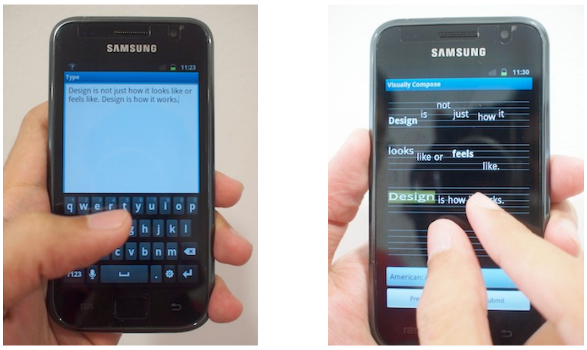


Figure 3. Illustration of the various Typography Mapping to Prosodic Information

SpeechPlay Mobile Application

SpeechPlay system consists of an Android mobile phone application called SpeechPlay Mobile and a large interactive touch-enabled display called SpeechPlay Surface. Users running the SpeechPlay Mobile application (Android 2.3 and above) can create a message either by typing a new message on their device or by copy-pasting an existing message from the SpeechPlay Surface (Figure 4a). Users can then augment the text (and hear the resulting synthesized voice) by performing different functional gestures (Figure 4b). For example, moving words to various positions by a single-finger tap-and-drag will modify the pitch (vertical displacement) and pause duration (horizontal displacement), stretching or compressing a word by a two-finger pinch gesture will change the duration of the spoken word. Double tapping on a word gives it a bold typeface that produces an emphasis on the audio output (volume increase). Based on the formatting and position of words in the composition, the application generates a formatted Speech Synthesis Markup Language (SSML) message. To listen to the composition on their mobile device, users upload the SSML message to SpeechPlay Surface, which will interpret the message on its TTS engine and generate a corresponding audio file. The audio file is then streamed back to the SpeechPlay Mobile application for an audio preview.



(a) (b)

Figure 4: Interface of the SpeechPlay Mobile Platform

SpeechPlay Surface

SpeechPlay Surface takes form as a large touch screen display (Figure 5). This acts as a centralized system for synthesized speech generation and an interface for downloading and collating compositions to and from SpeechPlay Mobile application. At the heart of SpeechPlay Surface is a computer (Macbook Pro, OS 10.7) running a TTS engine provided by CereProc (*www.cereproc.com*), which process the SSML messages from SpeechPlay Mobile to generate a synthesized output.

Compositions uploaded to SpeechPlay Surface will appear as an entry in an animated 3D tag cloud. A person standing in front of the touch screen can browse through the tag cloud and tap on a keyword to bring out its corresponding composition, which will be displayed on the right side of the screen beside the tag cloud. The person can then listen to the composition and choose to download the entire composition to their SpeechPlay Mobile application where they can add a new line, change words and/or re-augment the text. The modified composition can be shared back in the SpeechPlay Surface as a new entry.

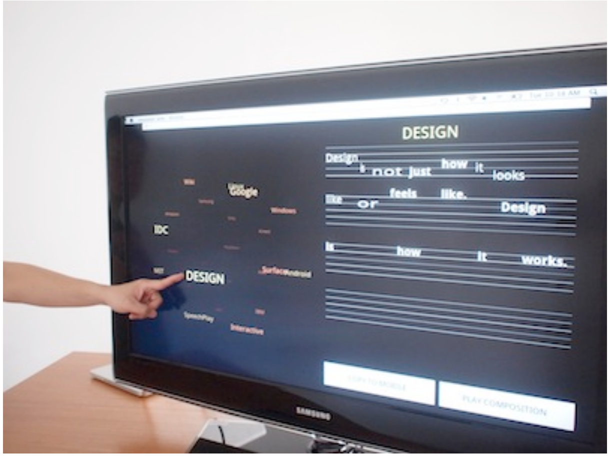


Figure 5: SpeechPlay Surface

USER REACTIONS

During the initial interactions with six users, we were able to identify two distinct user behaviors. Some users took a fairly long period of time to augment the message and were interested in getting a more natural and expressive speech output. The other users took the task in the opposite direction and came up with fun and often unnatural sounding speech output within a very short period of time. Most users were quick to pick up on the idea of manipulating words through gestures on their mobile devices and all felt that the relationship between the text typography and speech prosodic information was intuitive. Most users like the fun aspect of the system and the ability to add expressiveness to a plain text message. It was interesting to note that several users hoped for a musical component to the system where one could compose lyrics using the same interface given some background music as a guide.

FUTURE DIRECTIONS

SpeechPlay is still very much a work in progress. We are in the process of extending the existing work to allow augmentation of text at the syllable level within a word. This will open the possibility of adding more variation to expressive speech output. With the ability to control synthesized speech at the syllable level, we also plan to introduce a musical component to a future version of SpeechPlay. This will allow music and lyrics composition to be fun and interactive, making it a useful tool for composers. We also intend to look into applications of SpeechPlay for speech-impaired people.

ACKNOWLEDGMENTS

This work was supported by the International Design Center (IDC) of the Singapore University of Technology and Design with IDC grants IDG31100104A and IDD41100102A.

REFERENCES

[1] Ananny, M., Strohecker, C. and Biddick, K. Shifting Scales on Common Ground: Developing Personal Expressions and Public Opinions. *IJCEELL*, 14 (6), (2004), 484-505.

[2] Ballagas, R., Rohs, M., and Sheridan, J. G. Sweep and point and shoot: phonecam-based interactions for large public displays. In *Proc. CHI'05,* (2005), 1200-1203.

[3] Blankinship, E. and Beckwith, R. Tools for expressive text-to-speech markup. In *Proc. UIST 2001*, ACM Press (2001), 159-160.

[4] Breazeal, C. Emotive qualities in lip-synchronized robot speech. *Advanced Robotics*, (2003), 97-113.

[5] Cheok, A. D., Fernando, O. N. N., Wijesena, I. J. P., Mustafa, A., Barthoff, A. and Tosa, N. BlogWall: a new paradigm of artistic public mobile communication. In *Proc. Mobile HCI'07*, (2007), 333-334.

[6] Cytowic, R. E. *Synaesthesia: a Union of the Senses.* New York: Springer-Verlag, 1989.

[7] d'Alessandro, N., Pritchard, R., Wang, J. and Fels, S. Ubiquitous voice synthesis: interactive manipulation of speech and singing on mobile distributed platforms. *Ext. Abstracts CHI 2011*, ACM Press (2011), 335-340.

[8] Edwards, A. D. N. *Speech Synthesis: Technology for disabled people*. London: Paul Chapman, 1991.

[9] Fels, S., Pritchard, R. and Lenters, A. ForTouch: A wearable digital ventriloquized actor. In *Proc. NIME 2009*, 274-275.

[10] Holopainen, J., Lucero, A., Saarenpää, H., Nummenmaa, T., Ali, A. E. and Jokela, T. Social and privacy aspects of a system for collaborative public expression. In *Proc. ACE 2011*, ACM Press (2011), 1-8.

[11] Howard, I. P. *Human Spatial Orientation*. London, England: Wiley, 1966.

[12] Jacucci, G., Oulasvirta, A., Ilmonen, T., Evans, J. and Salovaara, A. CoMedia: mobile group media for active spectatorship. In *Proc. CHI 2007*, ACM Press (2007), 1273-1282.

[13] Jayant, N. S. Speech coding and text-to-speech synthesis. In *Proc*. *AT&T Speech Processing Symosiump* (1990).

[14] Lederman, S. J. and Klatzky, R. L. Haptic perception: a tutorial. *Attention Perception and Psychophysics*, 7 (2009), 1439-1459.

[15] Martin, K., Penn, A. and Gavin, L. Engaging with a situated display via picture messaging. In *Proc. CHI'06*, (2006), 1079-1084.

[16] McGurk, H. and MacDonald, J. Hearing lips and seeing voices. *Nature*, 264, 5588 (1976), 746-748.

[17] Peltonen, P., Salovaara, A., Jacucci, G., Ilmonen, T., Ardito, C., Saarikko, P. and Batra, V. Extending large-scale event participation with user-created mobile media on a public display. In *Proc. MUM 2007*, ACM Press (2007), 131-138.

[18] Rosenberger, T. Prosodic Font: The space between the spoken and the written. Masters thesis, Media Arts and Sciences, MIT, 1998.

[19] Scheible, J. and Ojala, T. MobiLenin combining a multitrack music video, personal mobile phones and a public display into multi-user interactive entertainment. *In Proc. Multimedia 2005,* ACM Press (2005), 199-208.

[20] Schröder, M. Expressive Speech Synthesis: Past, Present, and Possible Futures. *Affective information processing*, (2009), 111-126.

[21] Taylor, K. I., Moss, H. E., Stamatakis, E. A. and Tyler, L. K. Binding crossmodal object features in perirhinal cortex. *National Academy of Sciences*, 103, 21(2006), 8239-824.