

Expressive Text-to-Speech:

A user-centred approach to sound design in voice-enabled mobile applications

Peter Froehlich, Florian Hammer

Telecommunications Research Centre Vienna (FTW)
Vienna, Austria
(froehlich, hammer)@ftw.at

Abstract

Recent improvements in speech technology are expected to change the way we communicate, facilitating access to web services and applications (voice portals, multimodal email clients or games) in various mobile situations. In order to be attractive for users, however, speech output needs to be more expressive to increase naturalness, meaningfulness and ease of listening. Conventional Text-to-Speech (TTS) systems do not express non-verbal elements occurring in written text. This paper describes our user-centred approach to design non-speech sounds to express several non-linguistic elements in an email reader. Furthermore, it reports on a user study in which the effectiveness and satisfaction regarding the sound-enhanced emails was investigated. The results indicate that in most cases the addition of tailored non-speech sounds increases the user's ability to perceive the structure and meaning of the audio content and decreases the subjective workload, without leading to a higher user annoyance. As a consequence, non-speech sound should be integrated into market-ready TTS systems in a systematic way.

Keywords

Speech, multimodal, usability, non-speech sound

INTRODUCTION AND BACKGROUND

Speech technology is widely regarded as a promising alternative to overcome the input/output bottleneck in mobile computers (e.g. von Niman 2004). Mobile data services, such as voice portals and email readers offer unique features to users by providing access to personalised information on the Internet in mobile situations. Many research projects are working on solutions to provide mobile multimodal access to Internet applications (e.g. W3C's multimodality group). There are numerous combinations between input and output modalities, which can be chosen based on the given usage context.

When visually displayed text is converted to the auditory modality in such systems, non-linguistic elements and iconic symbols are difficult to be expressed by means of speech in a satisfactory way. In the case of a voice-enabled Email program, however, the auditory expression of the following elements would be a value-added feature to guide the user's attention and to improve the comprehension of the text content:

1. Elements helping to separate text into meaningful subgroups (bullet-points, separation lines, etc.),
2. Signs for hierarchical differentiation, e.g. indicating that a text passage has not been written by the sender (">>" etc.),
3. Highlighted text (by font formatting or capitals),
4. Emoticons, i.e. symbols based on convention that help to disambiguate and enrich text content (e.g. Smileys and sad Smileys, called "Frowneys").

Most current Text-to-Speech systems do not support the expression of this potentially valuable extra information is not expressed, although pre-processing rules could easily be generated.

A substantial amount of empirical research has shown that non-speech sound can be efficiently used to convey non-linguistic information in multimedia and audio-user interfaces (see Brewster 2003 for an overview). Several research projects have developed and evaluated concepts for transferring complex visual interfaces, such as GUI interface paradigms (e.g. Mynatt 1995) or web documents (James 1997, Asakawa et al 2002) to the auditory domain, in order to support visually impaired users.

When considering the adoption of non-speech sound in the competitive market of speech-enabled mobile applications (such as voice portals, multimodal games, or Email readers), neither these research results nor "common-sense guidelines" for multimodal interface design (e.g. Larson 2003) are sufficient. There is a lack of reliable empirical data about appropriate ways to combine speech and non-speech sound in market-ready voice-enabled applications and services. Especially the multi-faceted issue of user satisfaction (such as enjoyability and expressiveness, but also annoyance, quality of service, etc.) is critical for a successful adoption of such

features. Furthermore, the offered sound solutions will have to be as intuitive as possible, because - comparable to web usage (e.g. Nielsen 2000) - users will not accept a long learning phase for mobile applications and services. Empirically derived expertise in this area could encourage voice service providers to incorporate sound features in their services, and voice user interface designers could use these results as a reference for their work. Thus, our goal was to answer the following questions in a user study:

1. How can non-linguistic elements best be expressed within an Email reader?
2. To which extent can the expression of non-linguistic elements in an Email-reader support the user in acquiring and processing information?
3. How do users appreciate this additional expressiveness? Is it seen as valuable addition or is it regarded as source of annoyance?

This paper is structured as follows: In the next section, we will present our sound design approach and elaborate on our first research question. The following section describes the methodology and results of our user study which was designed to answer the second and third question. Finally, we briefly reflect on the chosen design approach, the user study results, practical implications, and future research issues.

CREATION OF SOUND EXPRESSIONS

Our approach was to take a theoretical reference frame (psychological grouping principles and theories of affective expression, respectively) as guidance for the creation of a relatively high number of prototypical sound expressions. For each of these sound expressions, we created prototypical audio emails by combining the sounds and the synthesised speech signal in a sequencer program. The audio mails were played back through a telephone handset to account for the special frequency characteristics in telephony. During the iterative selection and refinement, 5 expert listeners (audio engineers and natural language scientists) were providing qualitative user feedback.

In the following, the design decision rationale for each element is summarised:

Grouping: separation

As a design reference for auditory grouping, we considered Gestalt principles, which are still influential in visual interface design guidelines (Easterby, 1970), in auditory scene analysis (Bregman 2001), and in research on auditory grouping of musical sequences in cognitive psychology of music (see Snyder 2000). We transferred these principles to the specific interdependencies between speech and non-speech sound in TTS systems. In this context, the principles of similarity (different voices, different background sounds, non-speech sounds as a separation of voice streams), proximity (speech pauses, duration of separation sound, etc.) and common fate (common beginning and ending of speech and sound sequences) were most helpful to provide a framework for the creation of prototypical sound expressions.

Based on qualitative comments of the expert users, we came to the conclusion that short separation sounds are the most appropriate solution to separate text passages from each other. They were perceived as less obtrusive than using paused background sounds and more effective than simply using longer pauses between speech passages. Rising or falling pitch of tones for auditory bullet points or different voices in order to differentiate text passages were rated as inappropriate, because they created a false expectation of a qualitative difference between text parts.

Therefore, the final decision was to use a short “click”-like sound as a representative for separation elements (such as bullets):

(Sound example 1)

Grouping: hierarchical differentiation (e.g. citation)

Again we used the above-described Gestalt principles as a reference point for creating the sound expressions. The best valued alternatives were either to use a background sound or to include different speaker voices. Compared to simple separation sounds or pauses, both methods were best suited to differentiate text passages from each other.

Since different voices were planned to be used as a differentiator on a higher conceptual level (i.e. to differentiate between news and mail types, or the gender of the addresser), we decided to use background sounds as expression for a cited passage. Another reason was that different voices in an Email might lead to the user’s confusion and orientation problems. We are planning to investigate the optimal use of different voices in a follow-up study.

Background sounds for underlining text passages have also been investigated in other studies. Susini et al (2002) compared different versions of pink noise to express hyperlinks in web radio. In the context of mobile voice applications, however, noise would not be the most appropriate solution, because it might be interpreted as technical transmission problem or it might be overheard in noisy environments. The highest valued background sound alternative was a smooth pad chord (major add9) played back during a highlighted text passage, rather than single tones or melodies. Hence, this background sound was chosen to indicate citations.

(Sound example 2)

Highlighting

Of course, the most intuitive way to highlight text would be to realistically model human prosody. Therefore, we experimented with paralinguistic speech parameters (F0, spectrum energy, speaking rate, volume), using a tool recently developed for manipulating prosody (Pucher et al 2003). However, the possibilities of dynamically creating natural emphasis in modern speech synthesis systems are very limited. Posterior signal processing of recorded speech units to change the speech style results in a much lower signal quality (see also Black 2003). Furthermore, existing markup tags for Voice XML and TTS engines are not sophisticated enough to model prosody in a natural way (e.g. subtle F0 variations on the syllable level). Considering these practical and technological limitations, we decided to use non-speech sounds for the expression of emphasised text passages. Changes in signal volume were rated to be unsuitable and annoying, especially for noisy environments in which the volume dynamics should be small in order to maintain the usability of the service.

A common approach in research on auditory interfaces (e.g. in Asakawa 2002, James 1997) was to play a short sound before each highlighted passage (mainly an HTML heading). In our pre-tests, the expert listeners understood such sounds more as an expression for separation marks than as a highlighted text passage. In written Email texts, however, highlighted text is not restricted to headings, but is also used to stress certain sentences within a paragraph or words within a sentence.

As for the hierarchical differentiation, the most appropriate sound alternative was judged to be a background pad chord.

(Sound example 3)

Emoticons

In everyday email and chat communication, emoticons are used to increase expressiveness and to compensate the lack of nonverbal communication channels (such as facial expression or speech prosody) to convey various kinds of semantic and pragmatic meaning. According to literature in the field of internet psychology (e.g. Suler 2003), the most often occurring instances of emoticons are

- the Smiley :-), and less often, the Winkey ;-)
- the Frowney :-(

The meaning of these iconic representations can only be understood in combination with the concrete text semantics they are combined with (Walther and D'Addario 2001). Especially the smiley is used in different contexts:

- to express a friendly atmosphere
- to convey an ironic connotation to a certain text
- to amplify a jocular text content

Thus, a sound expression of a smiley would only be useful if it could serve in all these functions.

One possibility to express emoticons is to model the emotional connotation by a human sound. This could be highly intuitive, natural and appealing in many situations. However, in the light of the above-described ambiguity of contexts in which emoticons are used, this rather direct emotional representation could also be valued as inappropriate. As an alternative, a synthetic, more abstract musical representation might be better suitable in this respect, but less intuitive during first usage.

In order to investigate the difference between the human and the synthetic sound expression systematically, we decided to include both in the experiment. For the human sound expression, the concrete sound expression identified in the pre-tests was a "giggle" to convey a smiley and a sad and regretful sigh for the frowney.

(Sound examples 4 and 5)

The creation of synthetic sounds was inspired by psychological models of musical expression (Rösing 1993). For the smiley, a fast and lively, modulated rising tone sequence was rated highest, and the most appropriate expression for the frowney was a slow, monotonous downward bending tone.

(Sound examples 6 and 7)

METHOD

Sample

Due to the rather broad target population of voice-enabled mobile internet applications, we aimed at achieving a distribution over many different demographic groups. Eighteen (8 male, 10 female) paid, native-German

speakers took part in the study, which was conducted in individual 2-hour test sessions. The mean age was 31 years (min. 20 and max. 61), the professional status was varying (9 employees, 4 students, 4 freelancers, 1 retired). 3 persons had already used a voice portal or Email reader, none of them on a regular basis.

General Setup

All test parts had a within-subjects design, with the expression alternative as independent variable (mainly sound vs. speech-only). The specific subtest design is explained below. The order of the expression alternatives (e.g. sound vs. speech-only) and their combination with the text stimuli were varied in order to prevent systematic learning effects. The tests were conducted in our usability lab. The speech signal for the audio emails had been synthesised with currently leading TTS technology and mixed together with the respective sounds in a sequencer program. The audio mails were played back through a telephone handset to account for the special frequency characteristics in telephony. When test persons had to take notes, they were provided with headphones.

Procedure and Subtest Design

After a short introduction, participants who had not used an Email reader before were given a short demonstration of such a service. In order to get familiar with the meaning of the sound expressions and to investigate the user's first impressions, example mails for the 4 non-textual elements described above were played back. After this, subtests were conducted for each of the 4 main elements (see next sections).

Grouping (Separation): The subjects listened to emails containing bullet lists and were asked to take notes on a sheet of paper. They were instructed only to capture the most important pieces of information and to best possibly achieve a mapping with the structure of the original email text. Each text had about 200 words and 8-9 bullets, the number of sentences per bullet varied in each text from 1 to 4. There were 3 expression alternatives: (1) separation sounds between list items, (2) speech-only, but twice as long pauses between list items than between sentences, and (3) speech-only, equal pauses.

The participants' written notes were analysed afterwards regarding their consistency with the structure of the actual mail by counting the number of "grouping errors" and "omission errors". After each text, the participants specified the subjective workload imposed by the task by filling out the NASA task load index (TLX, NASA 1987). After all texts had been played back, the test persons specified which alternative they preferred.

In a second part, the test users listened to simple bulleted lists (1 short sentence per bullet) and were asked whether they preferred the sound version or the speech-only version. This part should answer the question whether sounds are perceived as helpful or rather annoying in situations when structure can easily be conveyed by other features than non-speech sound (semantics or speech pauses).

Grouping (Hierarchical differentiation): The test participants were provided with 3 printed email texts and were asked to imagine that they themselves had written them. They then listened to the answers of the fictive addressees via the email reader. They were instructed to take short notes of the respective answers. The answer emails contained citations of statements in the original mail. The expression alternatives were: (1) the cited part accompanied by a background sound, the normal text speech-only, (2) speech-only, longer pauses between cited part and normal text than between sentences, and (3) speech-only, no pause differences. After each of the text had been played, the test participants were asked to fill out the NASA TLX questionnaire. After all mails had been played, the test persons were interviewed concerning their impression (preference, annoyance, etc.).

Highlighting: The test participants listened to mails with news and press text content. There were 4 different text types: (a) short and simply structured texts, (b)-texts, in which only parts of a sentence were highlighted (c) more complex texts with different heading levels, and (d) long news texts. The expression alternatives were: (1) background sound during highlighted text passages and (2) speech-only. For text type (b), an additional expression alternative was to increase the speech volume during the highlighted text passage.

After having listened to the respective expression alternatives of each of the 4 text types, the participants were asked to specify their subjective preference and to give a rating on the annoyance of the sound expression.

Emoticons: The test persons were presented 8 audio emails containing emoticons (smileys and sad smileys, called "frowneys"). There were 4 text types, one for a frowney and 3 for the smiley. In order to account for the semantic diversity of texts with smileys, a systematic difference was made between friendly, ironical and joking texts. For each of the texts, the subjects listened to the following expression alternatives: (1) No expression, (2) synthetic sound, and (3) human sound expression. They then had to rank the alternatives with regard to their subjective preference.

RESULTS

Grouping: Separation

When taking notes, users made significantly less grouping and omission errors during the sound condition than during the two non-speech conditions. The difference between these 3 alternatives is highly significant

(Friedman-Test; Chi-Square= 20,0; p<.001). When comparing the alternatives pairwise by means of a Wilcoxon test, only the difference between the two speech-only alternatives was not significant. The workload score was significantly lower in the sound condition than in the other two alternatives (t-Test for paired samples, p<.01). The difference between the two speech-only conditions was not significant.

	<i>Mean number of grouping errors</i>	<i>Mean TLX score (min: 1; max: 20)</i>
<i>Sound</i>	1,4	10,6
<i>Different pause length</i>	3,1	12,4
<i>Equal pauses</i>	3,3	14,1

Table 1: Mean number of grouping (incl. omission) errors and the mean TLX score when taking notes for the 3 experimental conditions

Both for simple and complex texts, 15 of the 18 test participants preferred the version with the sounds. 3 persons preferred the longer pauses (while 12 other persons did not notice that there were different pause lengths). The separation-sounds were not experienced as annoying. The mean value on a 10-point rating scale (1="not at all annoying", and 10="very annoying") was 1,53 for the complex texts and 1,67 for the simple texts.

Grouping: Hierarchical differentiation

Listening to mails concerning citations, 15 test users preferred the sound alternative. 3 persons preferred the speech-only alternative with different pauses, because the background sound was annoying for them (see below).

There was a slight difference in the NASA task load score between the 3 conditions: the specified subjective workload was lower during the sound condition than in the other two conditions. However, this difference was not statistically significant.

The mean annoyance rating score for these sounds was 3,1 (SD=2,97). Although this mean value is not very high, there were some exceptions. 3 test persons explicitly found the background sounds annoying (rating scores of 7, 7, and 10).

Highlighting

The following table summarises the text person's subjective valuation concerning preference and annoyance:

<i>Text type</i>	<i>Alternatives</i>	<i>Preference N participants</i>	<i>Annoyance Mean (and SD)</i>
<i>1. Short and simply structured texts, highlighted headings</i>	Sound	15	2,7 (SD: 2,3)
	Speech-only	3	-
<i>2. Parts of a sentence highlighted</i>	Sound	9	4,2 (SD: 2,1)
	Volume difference	2	7,3 (SD: 2,4)
	Speech-Only	7	-
<i>3. Complex texts, highlighted headings</i>	Sound	16	2,4 (SD: 1,9)
	Speech-Only	2	-
<i>4. Long texts, highlighted headings</i>	Sound	16	1,83 (SD: 2,15)
	Speech-Only	2	-

Table 2: Overview of preference and annoyance ratings for highlighting in different text types

Especially in the case of headings (text type 1, 3, and 4), there seems to be a strong preference for a highlighting by sounds. As reasons for their preference rating, many participants said that they would better understand the structure of the text and that the attention would be guided to a higher extent. 2 persons did not like any kind of background sound (one of them was in general very sceptical to voice systems in general).

When regarding texts in which only parts of sentences are highlighted (Text type 2), it is clear that a simple change of loudness is not a good solution to express highlighting of parts within a sentence (annoyance rating of 7,3; rejection by most test persons). Although more than half of the participants preferred the background sound alternative, there were still 7 persons who favoured the speech-only alternative. Furthermore, the subjects' ability to recognise the highlighting function of the sound much lower in this text type than in the case of highlighted headings.

Emoticons

Concerning the preference rankings, a high diversity between the participants was observed. Users reported to have difficulties to give a general preference ranking due to the different semantic and contextual content. They would have preferred to have a choice of alternative sounds. However, a fact is that participants preferred the two sound alternatives in the 8 example texts significantly more often than the speech alternatives (Wilcoxon-test, values, $p < .05$). Also when asked for a final overall ranking, only 3 persons preferred the speech-only alternative, 8 preferred the human sound expression, and 7 the musical sound expression.

There was also a considerable difference in the preference rankings between the types. The human sound expression was most often judged as appropriate for the joking text type and for the texts containing a frowney, whereas it was judged least appropriate for friendly and ironical texts. The mean preference for the synthetic sound expression was more homogenous throughout the text types. Unsurprisingly, during first usage, the human sound expressions were more often recognised as sonification of emoticons than the synthetic sound expression.

DISCUSSION

Answering our general research questions, we can make the following statements:

First, we have shown that a research-driven design rationale together with an iterative development methodology can efficiently support valid sound expressions for specific user interface elements.

Second, the appropriate use of non-speech sound to convey non-linguistic information can result in a better usage performance and a higher user satisfaction. By the use of non-speech sounds, users can more appropriately structure textual audio information into chunks of information (by means of separation sounds). For certain kinds of text highlighting, background sounds can help to guide the user's attention and to reduce the cognitive load.

Third, the addition of non-speech sound is preferred to the speech-only presentations and does not result in a higher usage annoyance. Exceptions for this are the expression of highlighted parts within a sentence and some aspects of emoticons. Furthermore, the sonification of hierarchical differences in email texts – such as cited text – is helpful for users. Even the less clear results concerning the expression of emoticons indicate that most users appreciate the enrichment of speech synthesis by non-speech sound.

As a conclusion from these findings, the adoption of these features into voice-systems should be considered. In this regard, users should be enabled to control their preferred presentation style, because a minority of users may prefer the "text-only" variant while the other users favour the more expressive alternative. This diversity of preferences is especially important when considering auditory emoticons. For these emoticons, even a choice of different forms of musical and human sound expressions should be offered.

A reasonable solution would be to let users make configurations according to their personal preferences via a dedicated area in an accompanying web page. A short audio demo explaining the sound expressions during this configuration process could minimise the potential risk of irritating some users (especially in case of the emoticons). Apart from non-speech sound expressions, participants often expressed their wish to make configurations also for other features (e.g. voice type, speaking rate). A general customisation feature would therefore be a favourable alternative. These results should encourage more research to increase the expressiveness of speech-enabled applications and to integrate sound design into speech applications more systematically.

REFERENCES

- Asakawa, C., Hironobu, T., Shuichi, I., and Tohru, I. (2002), Auditory and Tactile Interfaces for Representing Visual Effects on the Web, in *Proceedings ASSETS 2002*, Edinburgh, Scotland.
- Black, A. (2003). Unit Selection and Emotional Speech, in *Proceedings EUROSPEECH 2003*, Genève, Swiss.
- Bregman, A. S. (2001). *Auditory Scene Analysis*. Cambridge MA: MIT Press.
- Brewster, S. (2003). "Non-speech Auditory Output" in J.A. Jacko and A. Sears (eds.) *The Human-Computer Interaction Handbook*. Lawrence Erlbaum Associates: Mahwah, NJ:
- Easterby, R. S. (1970). The perception of symbols for machine displays. *Ergonomics*, 13, 149-58.
- James, F. (1997). AHA: Audio HTML Access. *The Sixth International World Wide Web Conference*, (Apr. 1997), p.129-139.
- Larson, J. (2003). Commonsense Guidelines for Developing Multimodal User Interfaces. <http://www.larson-tech.com/MMGuide.html> (accessed 23 June, 2004)
- Mynatt, E. (1995). Transforming Graphical Interfaces into Auditory Interfaces. *Dissertation at Georgia Tech Institute*.

- NASA Human Performance Research Group. (1987). Task Load Index (NASA-TLX). *NASA Ames Research Centre*.
- Nielsen, J., and Tahir, M. (2001). Homepage Usability: 50 Websites Deconstructed. New Riders Publishing, Indianapolis
- Pucher M., Neubarth F., Rank E., Niklfeld G., Guan Q. (2003). Combining Non-uniform Unit Selection with Diphone Based Synthesis, Proc. Eurospeech, pp. 1329-1332, Sept. 2003, Geneva, Switzerland.
- Rösing, H. (1993). Musikalische Ausdrucksmodelle. In: Bruhn, H., Oerter, R., and Rösing, H. (ed.), *Musikpsychologie: Ein Handbuch*. Reinbek: Rowohlt, p. 579-587.
- Snyder, B. (2000). Music and Memory: An Introduction. Cambridge: MIT press.
- Suler, J. (2003). E-Mail Communication and Relationships, In: Suler, *The Psychology of Cyberspace*, www.rider.edu/suler/psycyber/basicfeat.html (accessed 24 June 2004).
- Von Niman, B. (2004). Mobile Communication: Simplifying the Complexity, *Business Briefings: Wireless Technology 2004*. URL: <http://www.bbriefings.com>, Accessed 11 June 2004.
- Walther, J. B. & D'Addario, K. P. (2001). The Impact of Emoticons on Message Interpretation in Computer-mediated Communication. *Social Science Computer Review*, 19 (3), 324-247.
- W3C Multimodal Interaction Group (MMI) website. <http://www.w3c.org/2002/mmi/> (accessed 24 June 2004).

ACKNOWLEDGEMENTS

We would like to thank SVOX ltd for providing their TTS technology for this user study. This work was funded by Kapsch Carrier-Com AG and Mobilkom Austria AG, together with the Austrian competence centre programme Kplus.