SCREAM: Screen-based navigation in voice messages

By Håkon W Lie, Per E Dybvik, and Jan Rygh



Abstract

The bitmapped colour screens commonly found on desktops is a powerful visualisation medium. The telephone, another common desktop apparatus, is on the other hand not very visual. New services offered through the phone system can benefit from using the visualisation power of the computer display. The SCREAM project creates a visual environment for navigating the data space of voice messages. Incoming voice messages are analysed, certain caller characteristics are extracted (e.g. gender), and the system renders and displays images that help the recipient navigate the messages.
This paper was presented at the 1993 IEEE Symposium on Visual Languages, August 25-27, Bergen, Norway.

1 Motivation

Networked computers have become an important medium for message handling. In many environments, text-based electronic mail is the primary choice for personal and group communication. Still, the telephone has a larger installed base, and will continue to be a part of our lives together with answering machines and voice mail systems. There are several reasons why a tighter integration between voice message systems and screen-based messaging systems are beneficial (6)(8)(9)(11). Having to check several mailboxes is inefficient, and the interface bandwidth of a touch-tone phone is limited. Sound is linear by nature and takes time to listen to, while a visual environment instantly can provide both overview of a number of items and more detailed information about a single item.

Today's personal workstations often include a high-resolution colour display which can be used to ease the navigation in the data space of voice messages. The goal of the SCREAM project is to create a rich visual environment for navigating voice messages. The recipient is presented with images that each represents a voice message, and together provide overview for selecting and sorting (i.e. navigating) messages.


The SCREAM project borrows its name from the work of art by Edvard Munch which, to most people, expresses an intense, high-pitched, hysterical scream from a female voice (figure 1). This is a highly advanced visualisation of an aural message, and - if presented to you by your answering machine - expresses several characteristics of the caller: it's a woman, she's screaming, and in a difficult emotional state.

In addition to gender and emotional state, the age, language, and accent are useful *caller characteristics*. The hypothesis of the SCREAM project is that these properties, if attractively conveyed to the recipient, will improve navigation in the data space of voice messages. One example: The authors know very few Swedish-speaking children, and if a message from a Swedish child comes in, we could make a good guess at the identity of the caller and perhaps the context of the message. There would be good reasons for treating that message different from a hysterical male Norwegian with a west-coast accent - of which we know many.


Figure 2 outlines the main modules of SCREAM and the data flow between them. The rest of this paper will describe the process in more detail. Also, the user interface of the application will be discussed.

2 Speech analysis

Speech can be analysed at different levels of abstractions. An audio signal - normally presented as a *wave form* can be transformed into the frequency domain and presented as a *spectrogram* . The spectrogram indicates how much energy the signal contains in the various frequencies as a function of time. The energy of an audio signal is perceived as loudness.

Voiced speech (speech is 'voiced' if the vocal cords are vibrating) consists of energy in a fundamental frequency and frequencies that are whole-number multiples of the fundamental frequency (3)(7). The fundamental frequency corresponds closely to the perceived pitch of speech. Typically, the pitch varies as one speaks - if not it's *monotone.* The fundamental frequency does not exist in unvoiced speech, and is even hard to detect in voiced speech (that is, for a machine).

Climbing the ladder of abstractions, speech recognition systems analyse the audio signal and output the textual representation of the utterance. Current speech recognition systems have limitations. Recognising a small number of isolated words spoken by one person is within reach of technology, but error rates go up as the number of speakers and words increase. Also, continuous speech is harder to recognise than discrete speech.

There is an unfilled gap between spectrograms and the textual representation, and this is where SCREAM finds its niche. Instead of just displaying a spectrogram, we try to elevate the visualisation by creating images that are more useful in a message context.

The first step in creating this process is to do a simple analysis of the speech signal. The results, which are passed onto the *caller characteristics* module are:

- the fundamental frequency of the audio signal
- the probability of having a fundamental frequency
- the energy of the audio signal.

3 Caller characteristics

Without knowing the identity of the caller, humans can determine, fairly accurately, a number of *caller characteristics* from listening to a voice message, e.g.:

- gender
- age
- emotional state
- language
- accent.

For a machine, the easier one to guess is probably gender/age, i.e. categorising the caller as *man, woman* or *child.* The fundamental frequency identifies these groups: male voices average around 100Hz, while the same numbers for women and children are 200Hz and 300Hz, respectively.

The emotional state of the sender is hard for a computer to detect. However, monotone speech is often perceived as sad, while the pitch of the excited caller typically will vary (1). The same is assumed to be true for the energy of the audio signal. Calculating the excitement factor from these variations is then straightforward.

We have not tried to determine the language or accent of the caller.

4 Rendering images

 What does sound look like? The question is rhetorical and no final answer can be given. However, there have been highly successful attempts to visualise sound. Munch's *Scream* has been mentioned, and the Disney animation *Fantasia* visualises pieces of music (figure 4). These are both works of art - a clear indication of where one is heading when rendering images depicting messages.

Some words exhibit relationships between the visual and aural. A voice can be described as 'dark' or 'light', and a colour can also be 'dark' or 'light'. The same phenomenon is found in languages other than English. Therefore, it is natural to choose a light colour as basis when the caller is a child. Consequently, the female voice is depicted using a darker colour, while the male voice is the darkest. It may also be useful to the recipient to see variations within each category - a bass voice e.g. looks darker than a tenor.

The *excitement* factor is a little harder to visualise, but language may once again give us hints. The term *colourful* can mean *full of variety or interest*. As described above, the excitement is calculated from the variations in the pitch and energy of the audio signal. Following this reasoning, colourful images should depict messages with a high excitement factor, while monotone speakers should get more monochrome images.

The *length* of the message does not qualify as a caller characteristic, but is still an important clue for navigation. Intuitively, physical size or length visualise the temporal length.

 Figure 5 shows the elements of our current rendering technique.

5 User navigation

The SCREAM project started out as an attempt to add another entry point to a multimedia mail system. The goal was to integrate the presentation of messages of different media into one coherent interface. The problem has since been simplified to only

present voice messages.

Auditive user interfaces found in most voice mail systems organise messages in linear or hierarchical structures. The user must build a quite complex mental model to get the overview of available messages.

 Bitmapped colour displays allow us to create a simpler model and give the user an instant overview of all messages. Images, each representing a message, are organised in an orderly manner, and provide users with direct access to the message of interest (figure 6). With limited interaction, users can navigate a large number of messages, and by clicking an image the corresponding sound file is played. The messages may be organised in several ways to fit the user's preferences using incoming time, priority, caller characteristics and length as sorting criteria.

The analysis and rendering mechanisms focus on creating one image that represents the whole message. Due to this, the project has concentrated on creating tools for navigating a number of messages instead of navigating within one message. The only element that resembles the time axis is a simple scroll bar that indicates time left when a message is played.

6 Future work

Some of the areas where we plan to focus our attention are:

- Interesting work has been done in the area of automatic*emphasis detection* (4)(2). By visualising emphasis, the recipient would better be able to navigate within a message.
- The images that are currently being rendered do not exhibit the artistic sophistication of our inspiration. The computer may never rival Munch, but improving our methods of rendering is certainly a priority.
- Dream analysis and the vocabulary of music are also candidates for future studies.
- Currently, the SCREAM environment only renders still images. Animation may better visualise temporal aspects of a message.

We also believe that the ideas we have investigated can be useful for other applications. For example, a compact disc player could improve user navigation by visualising the timbre, rhythm, and dynamics of the music.

7 Conclusion

Integrating telephones with a desktop computer has the potential of improving existent services. The SCREAM visual environment addresses the problem of navigating voice messages. By representing each message with an image that exhibits caller characteristics, we convey information to give the user an overview of current messages as well as more detailed information about each message.

The images representing messages borrow features from common associations between sound and images. The strength of these associations have not been tested through user experiments, and surely have room for improvements. Also, one may question the correctness of strengthening the association between e.g. the colour red and the female gender.

The user interface of the SCREAM project also lacks the affirmation of formal user testing. However, judging from using the system - and the frustration many people express when using regular voice mail systems - we believe the computer screen has the potential of becoming a preferred medium for voice message navigation.

Acknowledgements

The authors wish to thank Per Olav Heggtveit and Jon Emil Natvig for their support.

References

[1] Cahn, J E. The generation of affect in synthesised speech.*Journal of the American Voice I/O Society,* July, 1990.

[2] Chen, F R, Whitgott, M. The use of emphasis to automatically summarize a spoken discourse.*Proceedings of ICASSP'92,* 1992.

[3] Denes, P B, and Pinson, E N.*The speech chain: The physics and biology of spoken language.* Anchor, 1973.

[4] Hu, A C.*Automatic emphasis detection in fluent speech with transcription.* B. Sc. thesis, Cambridge, MA, MIT Press, 1987.

[5] Rabiner, L R, Scafer, R W.*Digital Processing of Speech Signals.* Englewood Cliffs, N. J., Prentice-Hall, 1978.

[6] Schmandt, C, Casner, S. Phonetool: Integrating telephones and workstations.*IEEE Global Telecommunications Conference,* 1989, 970-974.

[7] Seneff, S.*Pitch & spectral analysis of speech based on an auditory synchrony model.* Ph. D. thesis, Cambridge, MA, MIT Press, 1985.

[8] Stifelman, L J. Not just another voice mail system,*Proceedings of American Voice I/O Society Conference 1991,* 1991.

[9] Uhler, S. PhoneStation, moving the telephone onto the virtual desktop.*Proceedings of USENIX,* San Diego, January 1993, 131-140.

[10] Williams, C E, Stevens, K N. Emotions and speech: some acoustical correlates.*Journal of the Acoustical Society of America,* 52, 1238-1250, 1972.

[11] Zellweger, P T, Terry, D B, Swinehart, D C. An overview of the etherphone system and its applications.*Proceedings of the 2nd IEEE Conference on Computer Workstations,* Santa Clara, CA, 1988, 160-168.