

## Emotion is the Key to Intelligent Design: Making VUIs Natural and Expressive

**Author: Sheyla Militello**

The necessity of at least partial automation of customer support via CRM (mostly due to economic reasons), is anything but new. There are a huge number of companies that have considerable costs and resources dedicated to call/contact centres and that depend on them for a significant part of their business and support.

There is usually some conflict between the needs of people calling in, and providers that provide the service (see Ken Abbott, "Voice Enabling Web Applications", pp 88-104, 2001). Callers want to have their needs quickly and courteously met. From the end-user's perspective, nothing beats having an empowered and well-trained person on the organization's front-end. From the provider's perspective, well-trained and empowered customer service representatives are expensive. In the same way, providers want the call centre to please customers in a cost-effective manner. Users and providers do not use the same criteria when measuring VUI usefulness and effectiveness. The user's criteria are, for example, "Can I get the information or perform the transaction I want?"; "Is the result worth my effort to get it?"; "Do I feel like I'm receiving a valuable service?". While the provider's criteria are: "Does it reduce the load on customer service agents?"; "Are users satisfied with the experience?"; "Does it increase the number of users I connect with?".

By these criteria, a VUI that gracefully and elegantly ends up routing most calls to a human operator is not meeting providers' needs. On the other hand, a VUI that never routes a call to a human operator is not meeting the needs of some end-users. Achieving a balance between the sometimes-conflicting requirements of end-users and providers is part of the design process. Skewing the balance too heavily to the provider's side often results in a VUI that is overly comprehensive, loaded with options, and frustrating to use.

But is it really possible to automate a call centre without compromising on the user's needs? One significant finding emerged from a poll carried out by a market research company for the Market Validation "Vocal Browsing" project ([http://www.loquendo.com/en/company/international\\_cooperation\\_MV.htm](http://www.loquendo.com/en/company/international_cooperation_MV.htm)). Two interaction modes were compared - DTMF vs speech recognition - in an automated call centre architecture. One system comprised several DTMF menus many levels deep, and the other a VUI that routed the call automatically towards a category by recognising the caller's requirements, but then continued with sub-dialogues and requests for confirmations. Findings demonstrated that more than two thirds of consumers preferred vocal interaction. The reason given was that, even though the system was not completely free of errors, the response feedback received was felt to be "almost human".

The answer, therefore, is YES - *provided that* intelligent interfaces, voice recognition and other technological prerequisites are not automated to such an extent that they are NOT able to deal with the many human and emotional subtleties that a system of that type should be able to manage.

Assuming that all the technologies required are fully mature, therefore, which is the best type of interface for such a system? Much emphasis is being put on avatars and/or digital assistants ([http://www.loquendo.com/newsletter/newsletter\\_feb\\_2008.htm#clients](http://www.loquendo.com/newsletter/newsletter_feb_2008.htm#clients)). Are such solutions consistent with expressivity? Some argue that if an avatar is not able to communicate or understand emotions then it is NOT credible. Any software that has the "human face of an avatar" should ideally also be able to perceive or communicate to some extent the emotions of the person who is speaking, and the technology must therefore make inroads in this direction if virtual assistants are to have a lasting and credible role in human-machine interfaces.

The growing interest in VUI design is inspired in large part by the goal of supporting more natural human-computer interaction. However, this goal becomes particularly challenging when voice interfaces function in difficult user conditions. Applications that range from systems for education and training, to mobile usage in noisy environments, to transactions in natural living environments, are usually very complex, since they need to rely on effective natural language understanding, robust dialogue processing and context appropriate speech generation. Users' perceptions of the overall performance of VUIs can be greatly influenced by weakness in each of these different modules.

That effective error-handling techniques are crucial in developing VUIs is beyond doubt (see Deborah Dahl, "Practical Spoken Dialog System", Text Speech and Language Technology, Vol. 26, pp 41-63, 2005), however, some researchers have recently introduced the importance of the concept of the "listening behaviour" of a conversational agent. Actually, while human-to-human listening behaviour relies on a complex interplay of gestures and speech feedback, most conversational agents stay silent and passive while humans are speaking to them, such that the lack of a continuous visual and sound feedback may be unsettling. Actually, providing linguistic feedback may greatly reduce that uncomfortable sensation. Linguistic feedbacks are mechanisms which convey understanding, contact, and reactions to the communicative content. The ability of the component technologies (speech synthesis, dialogue modules, and the face of the agent) to provide linguistic and visual feedback is an emerging requirement of VUI design, since a more complete feedback may compensate for recognition errors and increase the usability of the voice interface.

A further emerging requirement deals with the appropriateness of the feedback with respect to the emotional content of the communication. While the recognition of emotions in speech is still a long-term research issue, text-to-speech technologies are beginning to include a variety of expressive cues that can be used in designing the interface when the basic emotional state can be detected on the basis of context analysis. A variety of VUIs and embodied conversational agents can, in fact, be developed for certain given types of users, for example for elderly people. In the latter case, if a system is being developed for facilitating the usage of the web by the elderly, it would be important that the interface be able to express encouragement - to persist despite of initial difficulties, for example. "Encouragement" is not a basic emotion, but an expressive feature that implies both awareness of the difficulty encountered, i.e. context-awareness, and appropriate voice characteristics, such as the right speaking rate, the appropriate intonation, and so on.

This and other related issues are currently being studied within the framework of European research projects, such as Companions ([www.companions.org](http://www.companions.org)). The problem of appropriateness of feedback is also present when advanced mobile service systems provide critical multimodal communication support for emergency teams during rescue operations, such as the application scenarios studied in the SHARE project during the past few years (2004-2007). Within the duration of the project the consortium developed a very advanced prototype of a mobile information and communication system for large scale rescue operations. This prototype (SHARE system) is a multi-user system for on-site cooperation which supports the work of fire fighting organizations in the field. The rescue teams can benefit from mobile and bi-directional communication based on 3G (UMTS) infrastructure and mobile WLAN networks located on-site during the emergency operation; the SHARE system allows to exchange structured multimodal information resources, including audio, video, text, graphic and location information. Vocal technologies have a fundamental role in multimodal interaction when hands-free interaction is necessary. In particular, the speech synthesis has to be robust enough to work under extreme conditions, where intelligibility has to be guaranteed. This requirement can be met only if the speech generation technology presents high pronunciation accuracy and appropriate speech fluency.

So, then, what about emotional content in VUIs for *The Design of Everyday Things*? "Emotion" is today one of the key words in the world of design. In the same approach as emotional design (see Donald Norman, "Emotional Design; why we love (or hate) everyday things", 2004) one recognises that previous conceptions of the interfaces and objects of everyday use, all geared towards functionality and usability, were limited and limiting: that is to say, we can not ignore the pleasure, or otherwise, that we get from objects that we use every day. That which each of us is, is also determined by the objects that we use: we choose them, we appreciate them not only for the function they perform for us, but also for the sensations that they give us. If pleasing objects perform tasks better, why shouldn't more agreeable artificial agents also perform better?

If the user experience of speech-enabled systems will increasingly take place between the telephone and the TV, leaving space for more complex person-system relationships, the realisation of an assistant that can become a personal companion is not so far into the future. In the case of the voice command Media Center (see SSN April 2007, pp 20), the choice has been made to equip the TTS with a portfolio of customisable voices, enriched with commonly used expressive phrases in order to achieve a consistent profile of the assistant's personality that can guide the user in managing their files in various media.

In conclusion, therefore, any use of text-to-speech should pay close attention to the expressivity and naturalness of the synthetic voice, and these expressive elements should, in turn, reflect the emotional state of the user as far as possible. An inventory of expressive cues as discourse markers can play an important role in such a process, improving the naturalness and expressivity of generated speech.

It is important to aim for natural-sounding speech in artificial agents, but it is equally important not to overdo this at the expense of intelligibility and fluency, which in some contexts is of greater importance than expressivity.

Artificial personal assistants will only play a useful role in the relationship between human and machine if the interface is built around the linguistic and emotional content of the human-assistant dialogue, taking into account such factors as politeness, humour, mood, etc.

Imagine finishing reading this rather tedious article and saying to your virtual assistant: "I'm tired. Play me some music." Wouldn't it be agreeable to hear a sigh of pleasure and an expressive voice replying: "Hey ...you're right! Let's relax with a little classical music!"?

Originally published in Speech Strategy News, go to: [www.tmaa.com](http://www.tmaa.com).

*Sheyla Militello is working in Loquendo in Marketing and Business Development area. She has been appointed as Project Manager for different EU founded project in Market Validation of Business Plan and Innovation Technology, including research and management with international partnership.*

*She graduated in Psychology in 1993 and was awarded a Master in Ergonomics at the Polytechnic of Turin.*

*She joined the Voice Solution and Professional Services department in Loquendo in 2001. She worked for CSELT (later Telecom Italia's research and development labs) from 1993, when she was in charge of User Interface Design and Usability assessment in the Voice Services and Applications department. She has over fifteen years' experience in the design and assessment of vocal interfaces and multimodal human computer interfaces.*