IBM Pervasive Computing
March 2003

Measuring Listener Impressions of Synthetic Speech

by Melanie D. Polkosky,
IBM Pervasive Computing

Contents

Executive Summary

Computer-generated, or synthetic speech has made significant advances in recent years. The goal of generating natural-sounding human speech is an elusive one, requiring the understanding of speech components as well as other factors such as the speaker's volume, pitch, mood and emphasis.

The measurement of how a person responds to synthetic speech – and how closely a synthetic speech system approximates the human voice – also involves several factors, including intelligibility and "naturalness". IBM's Mean Opinion Scale – Expanded (MOS-X) tool has proven successful in measuring – and improving – computer-generated speech.

This paper describes the speech components, technology advances, and measurement systems involved in synthetic speech technology. It was written for IT managers and staff who are evaluating synthetic speech systems.

Advances in Synthetic Speech Technology

An important aspect of speech technology development is determining how listeners will respond to computer-generated or synthetic speech. Over the past several years, synthetic speech has improved dramatically, and it can now be used to simulate the complexities of human emotion. But even a short phrase spoken in synthetic speech can sometimes cause a listener to wrinkle their nose and growl, "*What* is wrong with that voice?"

It seems so simple: combine sounds into words and words into sentences and eventually you have a talking computer, right? It's something the average person does all day, every day, with seemingly no effort at all. Of course, every now and then you get a little pause, your voice catches, or you get a cold and you sound a little funny. Those are the exceptions, not the rule. Somehow synthetic speech makes this simple thing seem so *difficult*. Why can't synthetic speech just sound *normal*?

The Components of Speech

The goal of achieving normal-sounding synthetic speech is a surprisingly complex one. Speech is a string of individual sounds, or phonemes, that are combined in a sequence. For example, the word "synthetic" is a string of eight separate sounds: [sInθεtIk]. Each sound in a sequence varies slightly depending on the sounds that precede and follow it; this phenomenon is called *coarticulation*. Some sounds, such as [n], can affect almost all of the other sounds in a short sequence, such as a word.

When words are combined into sentences, the relative loudness and pitch of each sound changes, based on the speaker's mood, what he wants to emphasize, and the type of sentence (consider the difference between a command and a question).

Even when sentences are combined into longer segments of speech, individual sentences change in loudness, pitch, and emphasis, adding an even greater variation to the individual sounds. So, the eight sounds in “synthetic” are slightly different depending on whether they are spoken individually, in the word “synthetic”, in the first sentence, or in a paragraph such as the one you are reading now. Even though the human brain easily adapts to these subtle differences and you hear the same word each time, the actual physical signals that reach your ear are quite unique.

In synthetic speech, coarticulation and overall loudness, emphasis, and pitch changes (known as *prosody*) are extremely difficult to recreate. Many of the differences between human and synthetic speech are due to coarticulation and prosody. In human speech, prosody reveals a great deal of information about the speaker’s personality, mood, what she wants to emphasize, and characteristics of the topic being spoken about. In synthetic speech, prosody and coarticulation are by-products of how individual sounds are sequenced, and are less likely to have the same loudness, pitch, rate, or emphasis as when the same words are spoken by a human. Given these differences between human and synthetic speech, how can we tell how synthetic speech sounds to the average listener?

The Pathology of Speech

Interestingly, human speech disorders can provide a great deal of guidance when it comes to measuring how people hear synthetic speech. The field of speech-language pathology is concerned with disorders of speech and language. Often, human disorders result in modifications to speech that can be compared to the “unusual” qualities of synthetic speech. Listeners’ impressions of a speaker with a speech impairment – like synthetic speech -- can be negative. Many of these speech disorders involve disrupted coarticulation and prosody, and these sound adjustments contribute to an impression of the speaker. Speech-language pathologists are clinically trained to describe speech impairments, and these same principles can be used to evaluate synthetic speech.

Measuring Speech Perception

Principles of speech pathology and psychology have been used at IBM to refine tools that measure how listeners perceive synthetic speech. In the past, researchers have evaluated only the intelligibility (or "understandability") and naturalness (or human-like quality) of synthetic speech. In the early days of synthetic speech, these measurements were enough to distinguish between relatively poor and high-quality synthetic voices. As synthetic speech has improved, most voices are now quite intelligible, but intelligibility measurements alone do not capture the real social impact of synthetic speech.

IBM has recently expanded on a measurement tool called the *Mean Opinion Scale (MOS).* The MOS is a questionnaire, which was originally used to measure the intelligibility and naturalness of synthetic speech. The *Mean Opinion Scale – Expanded (MOS-X)* measures a synthetic voice's intelligibility, naturalness, prosody, and social impression. As in speech pathology, it is important to measure a variety of speech characteristics to fully understand how a synthetic voice impacts its listeners.

Using MOS-X to Measure and Improve Synthetic Speech

The new MOS-X has recently been used to evaluate algorithmic and database changes in IBM's text-to-speech (TTS) voices. Samples of IBM's old female TTS voice and the new IBM female *SuperVoice*, which was released in August 2002, were provided to listeners, who rated them using the MOS-X scale. These ratings showed that the new voice resulted in substantial improvements to intelligibility, naturalness, prosody, and social impression, as compared with the previous voice (see *Figure 1*, which compares the previous female voice heard over the Web to the new female SuperVoice heard over both Web and telephone connections). These data confirm that IBM's efforts to improve its synthetic speech were highly successful – and that a sophisticated measurement tool can capture those improvements dramatically!

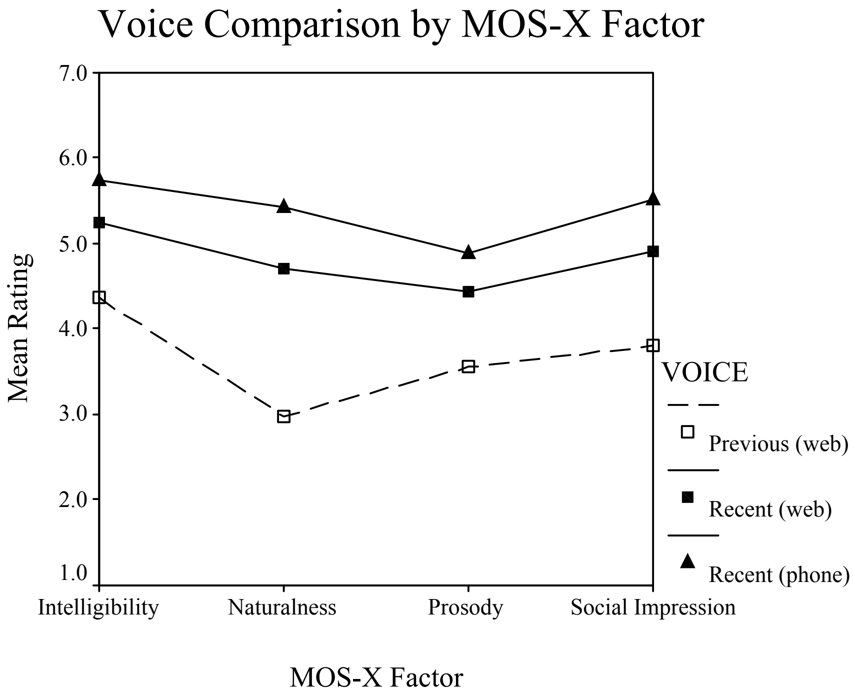

Figure 1 -- Voice Comparison by MOS-X Factor

Summary

While there is still work to be done in the advancement of speech technology, developing increasingly accurate measurement tools will be a vital part of that effort. Collaboration between human psychology and synthetic speech is an integral part of advancing these sophisticated forms of technology.

For more information, please see the article entitled "Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X" by Melanie D. Polkosky and James R. Lewis in the *International Journal of Speech Technology*.

References

Polkosky, M. & Lewis, J. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X, *International Journal of Speech Technology, 6,* 161-182.

This paper discusses strategy and plans which are subject to change because of IBM business and technical judgments.

All statements regarding IBM future direction or intent are subject to change or withdrawal without notice and represent goals and objectives only.

References in this publication to IBM products or services do not imply that IBM intends to make them available in any other countries.

Performance results obtained in other environments may vary significantly from your results. There is no guarantee that these measurements will be the same on your systems.

Trademarks

IBM Corporation
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
3-03
All Rights Reserved