Editor: Nevenka Dimitrova
Philips Research

Asynchronous Voice: A Personal Account

Keith W. Ross
Polytechnic University

Many forms of interactive electronic communication exist today but two forms are particularly pervasive. The first is synchronous voice, also known as the telephone, which has been around since the 19th century and ubiquitous for more than 50 years. Synchronous voice is pervasive in both our professional and nonprofessional lives. And with the widespread adoption of the cell phone, it has become overly pervasive for some of us—we'd prefer to eat in a restaurant in peace without being disturbed by a call or take in a movie without someone else's cell phone ringing in the theater!

The second pervasive form of electronic communication is asynchronous text, also known as electronic mail, which has been around for a relatively short time but we increasingly rely on it. Today, just about anyone who can read and write has at least one email account. Many of us also have a love–hate relationship with email—we love its features, convenience, and cost, but we lament devoting several working hours a day to reading and writing messages.

Love them or hate them, synchronous voice and asynchronous text are here to stay. They've become essential components of modern society.

Given the immense popularity of these two forms of interactive communication, one naturally wonders if they can be combined in some fashion to create a compelling new form of interactive communication. If we consider all variations of synchronous and asynchronous interactivity along with voice and text media, we obtain four combinations:

- *synchronous voice*: for example, wireline telephony, cellular telephony, Internet telephony, and ordinary face-to-face human conversation;
- *synchronous text*: for example, instant messaging;
- *asynchronous text*: for example, email and postal mail (text messages in cell phones are also largely asynchronous); and
- *asynchronous voice*: What the heck is that?

The first three combinations have led to killer applications. Yet we don't fully understand the last combination—asynchronous voice—which, in my opinion, remains to be fully explored. Before continuing, let's define asynchronous voice:

> Asynchronous voice is the interactive communication process of people leaving voice messages for other people and the other people responding with their voice messages.

A primitive form of asynchronous voice

A primitive form of asynchronous voice is a kind of telephone tag in which people use voice mail to have an interactive conversation. For example, Alice leaves a detailed voice mail message for Bob; Bob listens to the message and calls back, only to get Alice's voice mail; Bob then leaves his own detailed voice mail message; and so on. Voice mail, of course, wasn't intentionally designed for back-and-forth interactive conversation, although people do occasionally use it to communicate and collaborate in this manner. When people communicate through telephone-tag voice mail, it's usually not because they want to. In most cases, they'd prefer to talk live—that is, with synchronous voice.

So, the question I want to address in this article is, Are there compelling asynchronous voice appli-

1070-986X/03/$17.00 © 2003 IEEE

Published by the IEEE Computer Society

cations? I think so, and I hope to convince you.

How I got interested in asynchronous voice

Back in the mid-1990s—when the Web exploded on the Internet scene, when Netscape had Microsoft scared, and when streaming audio was only emerging—many academics including myself became interested in online education. I was especially interested in asynchronous learning, which was being espoused by Frank Mayadas, a program director at the Sloan Foundation. The Sloan Foundation provided some seed funding so that I, along with some of my colleagues at the University of Pennsylvania, could develop online asynchronous courses.

An asynchronous online course has two essential components: course content delivered over the Web and asynchronous interactivity, typically taking place in message boards (also known as forums and newsgroups), but sometimes using email distribution lists in conjunction with, or instead of, message boards. The asynchronous interactivity component is what really makes the course interesting—without it, the course is nothing more than a multimedia textbook delivered over the Web.

My own twist to asynchronous learning back in the mid-1990s was to deliver the content with streaming audio coupled to Web pages. There were surprisingly few academics experimenting with asynchronous voice at that time, even though streaming audio worked marvelously well over the then ubiquitous 28.8-Kbps modems. For the asynchronous interactivity, I used Web-based message boards extensively.

While preparing audio lectures and at the same time engaging in extensive asynchronous discussions with students over the message boards, I began to think about what would happen if streaming audio were combined with asynchronous dialogue. This led to the idea of a *voice message board*, in which messages are threaded and have text titles like an ordinary message board, but the messages themselves are voice messages rather than text messages. My feeling then (and now) was that a voice message board would be a powerful educational environment for a variety of subjects, including literature, poetry, public speaking, debate, and management—in fact, for any subject that has a significant amount of verbal exchange in the traditional face-to-face classroom setting.

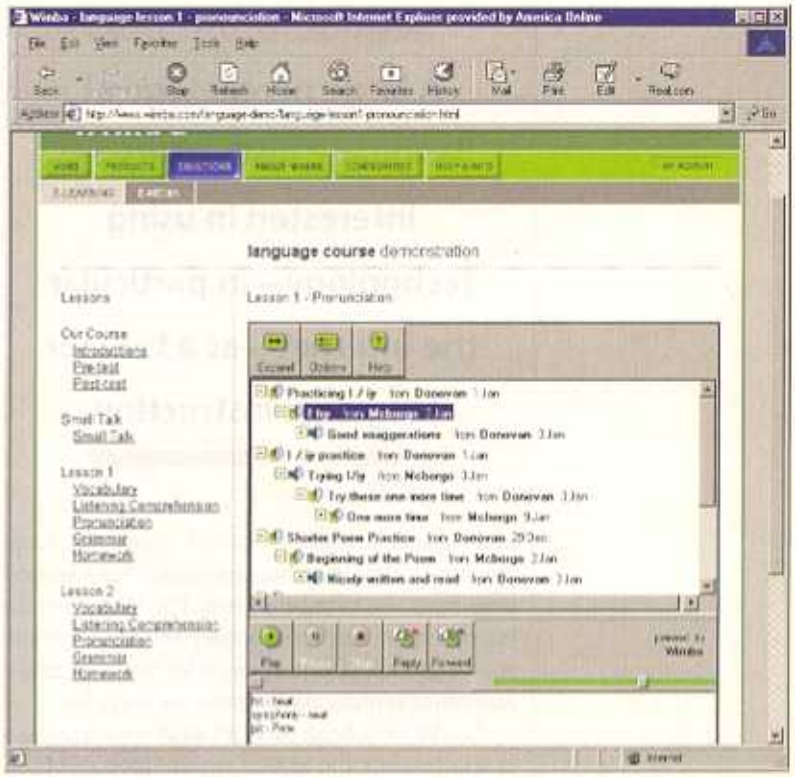

Figure 1. Voice board for pronunciation practice in an online English-as-a-second-language class.

A startup is born

Well, one prototype voice message board led to another,[1-3] and before I knew it, along with some PhD and master's students at the Institute Eurecom, we founded the startup Wimba (http://www.wimba.com) in the summer of 1999, with the goal of conquering the virgin asynchronous voice market. Wimba's first products were Web-based voice messaging boards and voice email. Figure 1 shows the GUI for the voice message board.

One feature of Wimba is that the client applications automatically download and install on the user's computer. Moreover, the clients run on almost all platforms, including Windows and Macintosh. The applications also run in the vast majority of installed browsers, including Microsoft, Netscape, and AOL browsers.

These features result from Wimba's patent-pending client-side architecture. This architecture consists of a Java applet and a small platform-independent agent program that performs audio capture and other audio tasks, including playback and automatic gain control. The applet is cached in the user's browser cache and the agent is cached in the user's file system. Because these programs are cached locally, during the second and subsequent visits to a Wimba application, the

> A large percentage of language instructors are interested in using technology—in particular the Internet—as a tool for language instruction.

application launches directly from the local machine rather than being downloaded a second time from the Wimba servers. The Wimba voice board application (Java applet and agent) is less than 150 Kbytes. The client sides for voice email and other Wimba applications are even less.

The Wimba design doesn't send voice messages as attachments, but instead streams messages from Wimba servers using a proprietary protocol that runs over the transmission control protocol (TCP). This streaming design renders the applications highly interactive and responsive, even for users who have dial-up modem connections. Because the stored messages are streamed over TCP, the client prefetches the message while the user listens to it. Typically, the client obtains the entire message—without any missing bits—well before the user finishes listening to the message.

Several standardized protocols for streaming media have emerged over the past few years, including real-time transport protocol (RTP), real-time control protocol (RTCP), and real-time streaming protocol (RTSP).[4] However, we soon discovered that many users are behind institutional firewalls, which often filter just about everything except HTTP. For this reason, many multimedia streaming products today use HTTP for client–server interaction, with the multimedia delivered in the HTTP messages. In particular, the Wimba server detects proxy firewalls and streams through the firewall over HTTP.

The Wimba technology not only streams messages from server to client, but it also uses upward streaming, which streams messages from client to server during message creation. Within a few seconds after the user finishes recording a message, the message has typically arrived at the server in its entirety. A user can record and send multiple voice emails or make multiple posts to a voice board without delays between messages.

Performing arts

Our original plan was to develop voice message boards and voice email products for e-learning, but like so many other startups in 1999, we got briefly sidetracked with developing community applications. In January 2000 we launched a trial community site, in which we featured voice message boards for a variety of topics, including jokes, poetry, romance, music, politics, and sports. After a little advertising, many people came to check it out. The whole experiment was a fascinating foray into the medium of asynchronous voice. Here are some things we learned:

- *Joke forum*: When delivered with conviction and skill, a vocalized joke can be highly entertaining, more so than the text-delivered jokes that circulate in email.
- *Poetry forum*: Many poets today use the Web to disseminate their poetry. A voice board lets poets vocally recite online their poems and get feedback from listeners.
- *Music forum*: Perhaps the biggest surprise was the immense success of the music forums, particularly for country, gospel, and rock music. Singers and musicians loved to express themselves in the voice boards and get asynchronous feedback from the Wimba listeners. This was surprising because the audio quality of the original Wimba message boards was only mediocre, comparable to that of a Global System for Mobile Communications (GSM) cell phone.

Online education: language instruction

After Wimba's brief sidetrack into community applications, it re-repositioned itself as an asynchronous voice company for online education. In the fall of 2000 Wimba began to market voice message boards and voice email products to online educators. A variety of academic disciplines became interested in the voice message boards and voice email, but one discipline in particular really latched onto it—foreign languages.

Language learning is a big business worldwide, generating tens of billions of dollars of revenue every year. Particularly big is so-called ESL, or English as a second language, not only in Asia, Europe, and South America, but also in the US.

There are millions of language instructors worldwide, and a surprisingly large percentage of them are interested in using technology—in particular the Internet—as a tool for language instruction. Indeed, the Internet is ideal for connecting pools of instructors (say, in the UK) with pools of students (say, in Japan) who are separated geographically.

Voice over IP is an essential part for any form of online language instruction. How can you learn a language without practice in listening comprehension and speaking? This leads to another important question. Should an online language course use synchronous voice or asynchronous voice? This issue is actually the subject of a heated debate at many of the computer-assisted language instruction conferences and workshops today. Some language educators argue that synchronous voice is preferable because it's what people actually use in the real world. Many other language instructors advocate using asynchronous voice, at least in the initial stages in the instruction cycle, for the following reasons:[5]

- For beginning students, asynchronous listening comprehension is significantly easier than synchronous listening comprehension because they can repeatedly listen to what the instructor has said.
- Similarly, after composing their own message, students can listen to it, compare it with a target phrase, and re-record it.
- As is always the case with asynchronous communication, the instructor and students don't all have to participate at the same time. This is a particularly attractive feature for online language instruction, where the instructor and student may be separated by time zones.

Probably the best approach for most online language students is to start with asynchronous voice (along with other standard grammatical and pronunciation tools) and, as the student progresses, use both asynchronous and synchronous voice for instruction.

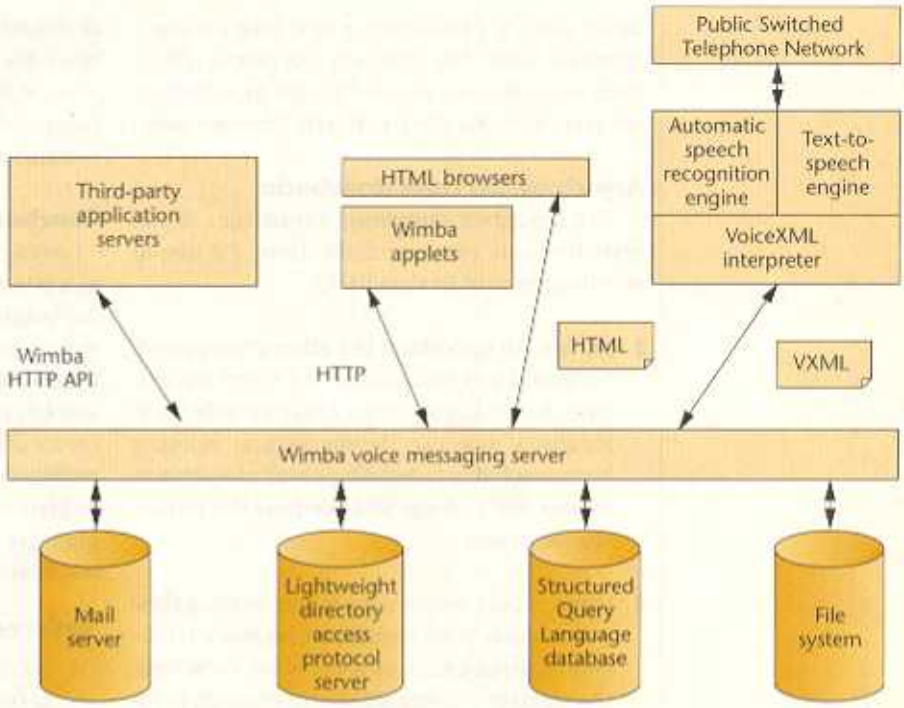

Figure 2. Accessing voice messaging applications from a phone or a Web browser.

Beyond entertainment and education

I've identified two broad contexts where asynchronous voice brings a clear added value. But will we ever use it extensively for communicating among friends or professional colleagues? Very possibly, but to do so, we must package it in such a way so that it's extremely easy to access and use. Access through a PC with speakers and a microphone isn't easy enough—PCs aren't always at our fingertips and microphones aren't ubiquitous peripherals. What we need is another way to access the message boards.

A natural access device for voice is, of course, the telephone (cell or wireline). We can use a phone to hear previously recorded messages and record our own messages. However, an ordinary phone doesn't provide a GUI, and without one, users can't navigate through a threaded message board. This is where speech recognition, text-to-speech, and VoiceXML come in. Text-to-speech lets phone users hear typed message titles, voice recognition lets users navigate through the threaded message structure, and VoiceXML lets us design and modify sophisticated voice response systems at a relatively low cost. Recognizing the importance of telephone access, Wimba has developed voice board and voice email products that users can access from both the Web and a phone (see Figure 2).

Interestingly, Navin Communications (http://

navin.com) is pioneering a new take on asynchronous voice. The company lets people reduce their long-distance phone charges by substituting asynchronous voice with synchronous voice.

Asynchronous voice drawbacks

I've described numerous advantages of the asynchronous voice medium. Here, I'd like to mention some of its drawbacks:

- *Privacy*: An important, but often underplayed, feature of text email is that it's silent and private. For example, in an office cubicle environment, you can devote several working hours each day to sending email messages to friends and nobody will overhear the private conversations.

- *Vanity*: Many people simply hate hearing their own voices. With asynchronous voice (voice boards, voice email, or some other variation), the message creators are always tempted to listen to their recordings before posting them. After hearing their voices, they might decide not to post the message or any voice message for that matter.

- *Monitoring*: Community sites that use text message boards often want to monitor messages to filter out postings that could potentially cause them legal difficulties. These community sites can use text-processing tools to tag undesirable messages. To do the same with voice requires powerful and costly voice-recognition tools.

- *Efficiency*: Readers can rapidly scan text messages to extract the information that interests them. With a voice message, listeners typically need to listen to the entire message to avoid missing something important. However, speech rate conversion technologies exist that let listeners fast-forward voice messages while maintaining speech intelligibility (for more information see http://www.nhk.or.jp/strl/publica/dayorinew/en/rd-0009e.html).

Although asynchronous voice has these drawbacks, synchronous voice suffers from many of them as well. Moreover, synchronous has other drawbacks that are particular to being synchronous (for example, requiring participants to converse at the same time and prohibiting the users from revising their words). Yet the telephone remains a killer application.

Conclusion

Asynchronous voice has already proven itself as a powerful medium for language instruction. Asynchronous voice also has great potential for enhancing existing forms—and creating new forms—of performing arts. But will asynchronous voice ever become an essential part of our daily communications? When a technology platform enables communicating with asynchronous voice to become as effortless as communicating by telephone or by email, asynchronous voice just might someday attain the coveted killer app status. **MM**

References

1. S. Cho and S. Carey, "Increasing Korean Oral Fluency Using an Electronic Bulletin Board and Wimba Based Voiced Chat," *The Korean Language in America*, vol. 6, 2001, pp. 115-128.
2. D. Turner and K.W. Ross, "Asynchronous Audio Conferencing on the Web," *Advances in Intelligent Computing and Multimedia Systems, Proc. Int'l Symp. Intelligent Media and Distance Education* (ISIMADE 99), Int'l Inst. for Advanced Studies in Systems Research and Cybernetics, 1999, http://ftp.csci.csusb.edu/turner/papers/aconf/aconf.doc.
3. D. Turner and K.W. Ross, "Continuous Media E-mail on the Internet: Infrastructure Inadequacies and a Sender-Side Solution," *IEEE Network*, vol. 14, no. 4, July/Aug. 2000, pp. 30-37.
4. D. Turner and K.W. Ross, "Comprehensive Architecture for Continuous Media E-mail on the Internet," *IEEE MultiMedia*, vol. 8, no. 2, April–June 2001, pp. 88-98.
5. J.F. Kurose and K.W. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, 2nd ed., Addison-Wesley, 2002.

Readers may contact Keith W. Ross at Polytechnic Univ., Dept. of Computer and Information Science, Six Metrotech, Brooklyn, NY 11201; ross@poly.edu.

Contact Visions and Views editor Nevenka Dimitrova at Philips Research, 345 Scarborough Rd., Briarcliff Manor, NY 10510; nevenka.dimitrova@philips.com.