

Comparative evaluation of voicemail-to-text services

January 18, 2010

William Meisel
President, TMA Associates (www.tmaa.com)
wmeisel@tmaa.com
(818) 708-0962

Introduction

Voicemail-to-text services are becoming increasingly popular. Transcribing the voice messages into text supports “visual voicemail,” which allows the messages to be reviewed quickly and dealt with in any order. The transcription of the voice files can often make it unnecessary to listen to the audio, but—even with errors—can also allow a better prioritization of messages than just the metadata associated with the message. The messages can also be archived as email, while it is not particularly useful or practical to archive voice messages.

Four voice-to-text services were tested for comparative accuracy. The services tested:

- Ditech’s PhoneTag;
- Microsoft’s Preview in Microsoft Exchange;
- Yap; and
- Google Voice voicemail-to-text.

All used the services used only speech recognition technology in this test —no human transcribers. Ditech was a partial sponsor of this research.

Methodology

The focus of this study is *comparative* accuracy, as opposed to *absolute* accuracy. While the results give an indication of absolute accuracy, the evaluation method, as discussed later, can give numbers that are pessimistic relative to typical performance that an individual might experience using the services.

Data

Services are compared with the same speech data. The data consisted of 500 messages, some short, some in excess of 45 words. The message content was derived from actual voicemail messages, with names and numbers changed where relevant. The Appendix contains examples of the messages used.

Five speakers were used, each speaking 100 different utterances typical of voicemail; each speaker spoke different messages. Thus, there were 500 utterances submitted to each service—the same utterances for all services. Two of the five speakers were female, and ages ranged from early 20s to over 60. There was a mix of landline and mobile calls. This is not a wide statistical sample, but represents typical variability in callers.

The 500 voice messages used in testing were recorded over the telephone channel. They were then placed as calls by a computer system to the services. Thus, every service received the same speech data.

Measuring word recognition accuracy

The US Defense Department's Defense Advanced Research Projects Agency (DARPA) sponsored a highly successful research initiative to develop speech recognition technology starting in the 1970s, with an annual meeting comparing results from different organizations using the same corpus of speech data. This allowed objective evaluation of which technologies were delivering the best results, as opposed to papers reporting results on different speech data, with no means of comparison. (The DARPA competitions are considered by many to have led to the wide adoption of Hidden Markov Model technology, the core technology used by many speech recognition engines today.)

A single accuracy number was useful for comparison, and DARPA specified the accuracy measure. The measure compares two utterances by finding the optimal alignment of words actually spoken versus those recognized in an utterance, and reporting the sum of insertion errors (extra words), deletion errors (missing words), and substitution errors (a word recognized as another word). The details of the DARPA accuracy measurement algorithm are described in a number of sources, including *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Prentice-Hall, 2001, pp. 419-421. The methodology avoids any subjective element in the evaluation.

Absolute accuracy

The error rates, as previously noted, are pessimistic. They were computed with a measure that would, for example, count a single word that was misrecognized as two words (e.g., "Goldsmith" as "gold Smith") as two errors (an insertion error and a substitution error), although a person could consider this one error. (Two words recognized as one word would also generate two errors.) This measure can give an "error rate" greater than 100% on a specific utterance, since the denominator is the number of words spoken. For example, if a caller says, "Joan, call Frederick," and the transcription is "Joe, call Fred or Rick," the error measure is 4 errors out of 3 words spoken—133%. This aspect of the error measure thus raises the average error reported.

Some errors arise from spelling differences that are essentially equivalent from the viewpoint of the message recipient (e.g., "Christie," "Christy," and "Kristie"), but which a computer algorithm would consider an error based simply on matching words. Most of these equivalency issues were handled in the evaluation (e.g., "five" and "5", and "Dr." and "Doctor" are equivalent), but not all.

Further, the personalization of a voicemail box allows adaptation to minimize errors in actual deployments, e.g., knowing the mailbox owner's name ("Hi, Bill") or contacts that call frequently. Personalization was an option not available in this study since all the messages were delivered to the same mailbox irrespective of content.

Finally, the accuracy is not a direct indicator of whether the "gist" of the message is clear. Errors created by repetitions or hesitations by the person leaving the message wouldn't necessarily impact the recipient's understanding. One general observation is that all the services did well in recognizing telephone numbers accurately.

On the other hand, the accuracy numbers are in fact a percentage word error number that gives a rough feel (perhaps an upper bound) for perceived accuracy.

Results

The results of the test are shown in the following table. The average error is the DARPA measure, as described. The rank is based on the relative score on each utterance, that is, the best-scoring service on the utterance was given rank 1. Ditech's PhoneTag was the most accurate at 14.2% word error and ranked highest on average.

	PhoneTag	Google	Yap	Microsoft
Average error*	14.2	17.7	21.5	16.1
Average rank**	2.0	2.2	2.5	2.1

* Expressed as a percentage by adding all insertion, deletion, and substitution errors

** 1 is best rank, ties were given equal ranking

Appendix: Examples of voicemail messages used in the test

Hello Amber, How are you doing? It's Evelyn. If you can, give me a call at your earliest convenience. Telephone number is 942-395-6310 942-395-6310. Thanks. Bye now.

Hi, this is Joanne from Dr. Hunter's office, Susan, at (612) 642-9210. This is the second time I'm calling. I'm sorry I missed your call, but, at your convenience, please call the office. Thank you, bye bye.

Hey Jasmine, it's Ed. I got your voicemail today. I just completely forgot to call you back. I apologize, but I'm at home (213) 861-5255.

Hey, it's me. I'm calling you back. Sorry I missed you. I'm close the phone, so give me a call. Love you. Bye.

To forward this call to another phone number, Press 2. To repeat this menu, press the pound key. I'm sorry, I did not hear you.

Hey Julie, please contact our office at 1 (866) 624-0146. Again, that number is 1 (866) 624-0146.

Hey Ian, it's Amanda. I just want to know if you have your phone charger. If not, I think I have it here, but I'm not sure if it's yours or not. So give me a callback. Bye.

Hey, Luke. It's Marilyn. Call me. Thanks.

Andy, it's Nathan. Call me on my mobile please. Thanks.

Hey, call me please.

Hey, I wanted to connect on Savanna and Cathy O'Hara and the other thing, so call when you can. Bye.

Hi Lucy, this is Carrie. It's about 4:15 – 4:20 or so on Wednesday. Emily just called and wants us to get on the phone together to chat with her, so you can give me a call at 714-283-9282 714-283-9282. Thank you, thanks so much. Bye.

Hello, this is Scarlett Ford returning your call again about computer repair. This isn't gonna work if you don't answer your phone. I will be home this evening after 7 o'clock if you wanna call me at 202-689-0165. Hope you're great! Bye.

Hi, this is Joanne from Dr. Fields's office, Shiela, at (622) 642-9210. This is the second time I'm calling. I'm sorry I missed your call, but, at your convenience, please call the office. Thank you, bye bye.