A study on speech over the telephone and aging

Maxine Eskenazi and Alan W Black

Language Technologies Institute, 5000 Forbes Ave.,
Carnegie Mellon University, Pittsburgh. PA. 15243 USA
max@cs.cmu.edu, awb@cs.cmu.edu

Abstract

We describe an experiment to show how the comprehensibility of speech over the telephone is related to the age of the listener. Our intention is to show figures to prove the commonly-held belief that as we get older our hearing of information over the telephone degrades. The study was set up to determine, for all age groups from 20-29 to 80-89, whether comprehension degrades with age and with the type of speech (synthetic or natural). We gave subjects sentences containing target word pairs that they were to write down. The pairs contained more or less predictable words.

Our findings, which we consider to be preliminary due to the sample size, show degradation of comprehension with age and degradation from natural speech to synthetic speech.

1. Introduction

This paper describes an experiment to show how the comprehensibility of speech over the telephone may be related to the age of the listener. Although it is believed by many that as we age, our ability to understand natural or synthesized spoken language, over a telephone degrades, most evidence is anecdotal at best. We therefore devised and carried out the following simple experiment to serve as a point of departure for examining and proving the degradation. The term, ”understanding”, is defined here as hearing, processing, and correctly reacting to spoken computer output since one or more of these three activities may be degraded for any given elderly user.

Our motivation for confirming this hypothesis is to fuel research oriented toward improving the usefulness of telephone-based information presentation systems for elderly users. The elderly do not for the most part have computers to access the internet and the information that they need for everyday life. They do, however, have telephones and, if dialogues with the elderly users are conducted correctly, the phone can replace the computer as a means of access to information such as bus schedules and medical appointments. A dialogue system could also be used in such applications as the automated home nursing robot for the elderly [2].

Using this newly-gained understanding of how speech comprehension degrades as users age, we can accordingly improve speech synthesis to be clearer to the user and oral dialogue quality to be efficient and effective.

2. Experimental Design

The goal of this experiment was to determine whether understanding of the content of an oral message heard over the telephone lessened as people aged and as the source of the speech went from natural speech to synthetic. The design of the experiment had to incorporate uncomplicated instructions so that elderly subjects would have no problem understanding what they were to do. It also needed to be short so that those who might be frustrated that they did not hear a token would not give up before the end of the test. We were specifically interested in understandability rather than other aspects of usability and therefore only concentrated on a limited number of variables. The basic scenario was to play several simple utterances, some naturally spoken by a human and some synthesized by a speech synthesis system, over the phone and to have users write down what they heard.

Evaluation of synthetic speech is a nortorious difficult problem, [3]. Subjects' views of synthesis quality are easily influenced by order of presentation, familiarity with synthesized speech and how questions are asked about their perception of quality. Naturalness and understandability can be independent. This work is just the start of a longer project to improve the quality of speech synthesis in general by focussing on a class of users who are known to have particular difficulty in understanding synthetic speech. In this experiment we decided to use standard diphone speech rather than more recent unit selection techniques which can often produce speech indisguishable from recorded prompts especially in limited domains. With diphone quality speech we also have the opportunity to impose predicted prosody.

Four different voices were used to deliver the utterances.

NN : was a natural spoken utterance by a female speaker

NS : was a natural spoken utterance by the same female speaker but after she was told that the listener couldn't understand what was being said, thus the utterance, was slower, more articulated and partly ”shouted”.

SN : was synthesized using a standard ”diphone” female synthesis voice.

SS : was again synthesized by the same standard ”diphone” voice, but unlike SN, where the prosody was predicted by a statistical model, here the natural durations and F0 were extracted from NS and imposed on the synthetic segmental form.

For each voice two utterances were presented, the first was of the form:

Please write down the following time ...

while the second utterance was of the form

Please write down the following words ...

The first utterance was intended to be much easier to understand due to the fact that the set of possible times is much more constrained than the second set of possible words. We deliberately chose to play the more predictable utterance first so the human listener could become more accustomed to the voice (be it natural or synthetic).

The sets of words consisted of pairs of relatively hi-frequency bi-grams, in all only four different pairs were used, *rose bush*, *Rose Bowl*, *holiday season* and *holiday shopping*. In designing the prompts we aimed at a medium level of difficulty. We thus avoided selecting more confusable word pairs, for example pairs which were one phone away from more common bi-grams, e.g. *baseball pat*. Although we were prepared to change the lexical content and difficulty to ensure the right level of difficulty for our audience, our pre-testing showed that our first selection was adequate.

3. Experiment

Each subject was presented with 8 utterances, and was asked to write down what she heard. They heard 2 utterances (one "time" and one "words") from each of the 4 voice types. The order of the voice types was changed for each subject (24 orders in all).

The system ran on a Linux machine using an inexpensive LineJack telephony card. A custom-written program was used to run the experiment.

The natural utterances were spoken by a graduate student who has experience in delivering prompts for telephone systems, and therefore can consistently deliver different utterances in the same style. The synthetic utterances were produced by the Festival Speech Synthesis System [1], using the `us1_mbrola` voice, considered by many to be a good high-quality diphone voice. Although we are aware of other synthetic voices which may have better understandability and/or naturalness, at this stage we were not concerned with measuring the quality of different synthesized voices, so we selected a typical example of easily achievable synthesis quality which is also publicly available.

The experiment was set up so that when a call was initiated, a two-key index was requested to select in which order the voices would be presented. The experimenter normally did this for the subject so as to avoid lengthy explanations of the experimental setup. Once the order was selected a single key press allowed the user to hear the first and following utterances. In this test we did not allow the user to hear the utterance more than once. The possibility of being distracted from hearing an item by such sources as asking questions, abrupt noises, and other background noises was thus kept to a minimum. Users had no problem learning this method; we only eliminated one of 536 responses due to distraction. Even when the user had a telephone where the keypad was in the handset, the system response time after the user pushed a key was long enough that the subjects did not miss any part of the utterances.

The subjects were given a sheet on which they were to fill out their initials, age, whether they had hearing problems, and they signed a release statement. On the bottom half of the sheet were the 8 response lines, preceded by the instructions. The subjects were asked to read the instructions and the instructions were also read out loud to them by the experimenter (eight utterances, asking them to write something down each time, press a key other than pound sign to go on). The font on the sheet was at least 16 points in every part.

4. Subjects

Carnegie Mellon's Homecoming event has, in past years, provided us with much useful data. Alumni willingly participate in experiments and we obtain a more general user population than by eliciting participation just from students and colleagues. Since Homecoming attracts many older alumni, especially those who are there for their 50th reunion (age of about 70), we felt it would be the perfect place to carry out our study.

We set up 2 telephones close to the registration area and found that many people volunteered for the experiment readily. Many had strong opinions on the problems that reduced hearing/understanding, above all other forms of perception, creates in their lives.

There were 67 subjects in all, with ages ranging from 24 to 87. We grouped them by ten years of age (20-29, 30-39, etc.) in our results. The age groups above 50 years old were more populated than the groups under that age.

5. Results

The figures below show the results of the study. There were 67 subjects in all and they each had a total of 8 utterances to listen to. This gives us a total of 536 tokens. From this, as mentioned above, there was one response that was eliminated due to a distraction.

Figure 1 shows how, over all of the conditions combined, hearing decreases with age (decrease being defined as a lower percentage of correct scores). The x axis represents the age groups and the y axis represents the percent correct. We can see that as people get older, their hearing in general decreases.

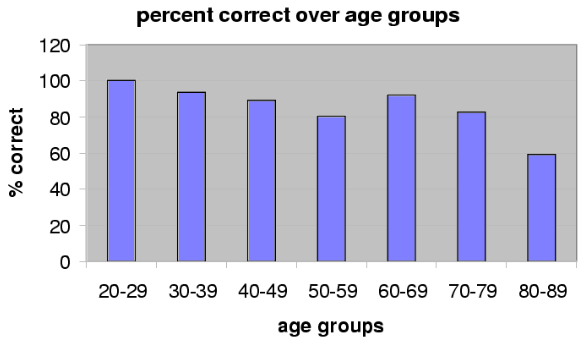

Figure 1. Percent correct responses per age group for all conditions combined

We then looked at the results for the eight different categories over all speakers. This is shown on Figure 2. The x axis represents the different conditions and the y axis again represents percent correct responses. We have shaded the natural speech bars in a light color and the synthetic speech in a darker color.

We can see here that synthetic speech was in general harder to understand than natural speech. We can also see that "time", with its reduced semantic and lexical possibilities, was correct more often than "words" within the same condition. We note that the recorded speaker's attempts to be better understood in the NS conditions did not meet with better success in general than the natural speech NN condition. This may be because this speaker chose to speak much louder, instead of inserting pauses, or increasing F0 emphasis. It is well known that increasing amplitude does not make speech much easier to understand for older listeners. Their problems concern attention, thus such elements as pauses and F0 placement serve to attract their attention to specific high-content words.

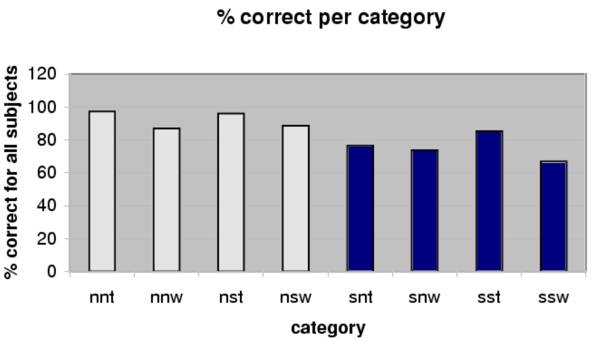

Figure 2, Results of all speakers over the eight different conditions (NN = natural speech, NS = natural very clear speech, SN = synthetic diphone speech, SS = diphone speech with natural F0 contours; t="time", w="words").

We also looked at more specific aspects of the data shown in the above two figures. Figure 3 shows a breakdown of the results with different conditions (categories) as a function of the age groups. We see that while understanding of synthetic speech rapidly decreases, natural speech remains understandable much longer.

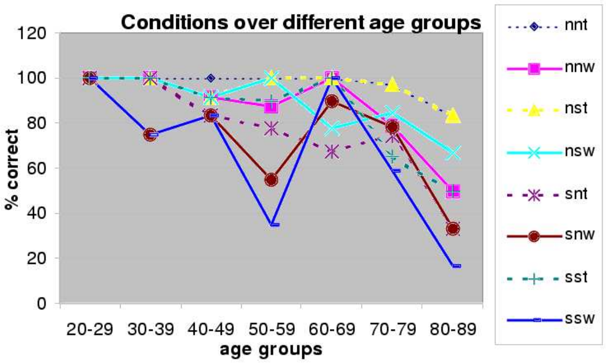

Figure 3. Comparison of different conditions over the age groups.

We also broke results down in Figure 4 to show the age groups as a function of the different conditions. We note that in almost all cases, as expected, times are easier to understand than words. However, for the SN case of purely synthetic speech, this is not always true. This implies that natural prosody actively helps people to understand speech. It would also seem, from our data, that at about the age of 80, comprehension declines greatly over all conditions

The correlation between age and percentage correct is -0.33, that is, the older the person, the less they understand.

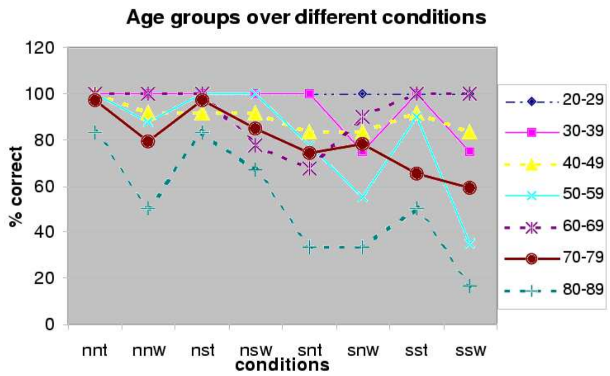

Figure 4. Comparison of different age groups ovre the different conditions.

6. Discussion

Since our results are based on only 67 subjects and 8 tokens, we cannot consider them to be decisive - more data would be needed for that. But, they do show us trends that are, in our belief, very important. They show that speech synthesis is not a solved problem. As individuals get older and rely more on the telephone, that lifeline is effectively not serving them as it should. They understand speech that is less natural, but the synthetic speech shows a pronounced decrease in understanding.

We also note that some information is easier to understand than others; it would seem that our ability to predict items in a semantically restrained domain remains constant as we age.

We have anecdotal evidence about hearing problems. In general, people who admitted they had hearing problems did better then those who did not (but whose spouses had dragged them over to our experiment saying they really needed to participate, implying their hearing was not very good). This implies that those who admit they have a problem and do something about it will continue to understand better than the others.

We should also note that the people who attend Carnegie Mellon Homecoming are a specific group. They have obtained higher education and are very mobile. It is evident that the largest part of the elderly population who desperately need to use the telephone for essential services are not as highly educated and, especially, are not as mobile (would not have physically have been able to attend Homecoming). We therefore need to extend and validate our data by repeating the experiment in activity centers and assisted living communities for the elderly.

Finally, we can see that the clues that our speaker used to try to be better understood were not completely successful. But we can measure the elements of prosody (timing, intonation and phrasing) that she employed and conduct further studies that use other combinations of these elements in order to find which elements make synthetic speech more understandable.

7. Conclusion

From this preliminary study, we conclude that speech synthesis over the telephone is not a solved problem for an increasingly important part of our population and that research efforts need to be concentrated on changing this.

From this preliminary study, we can already see that there is a decrease for all types of speech over the telephone channel from 100% comprehension at ages 20-29 to just under 60% at ages 80-89. And, clearly, the synthetic speech we used is less understandable than natural speech. These findings imply that more natural synthesis quality is necessary to improve understanding.

8. References

[1] Black, A., Taylor, P., and Caley, R. The Festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival.html, 1998.

[2] Roy, M., Pineau, J., and Thrun, S. Spoken dialogue management using probabilistic reasoning. In *Proceedings of 38th ACL* (Hong Kong, 2000).

[3] van Santen J., Pols, L., Abe, M., Kahn, D., Keller, E., and Vonwiller, J. Report on the third ESCA TTS workshop evaluation procedure. In *3rd ESCA Workshop on Speech Synthesis* (Jenolan Caves, Australia., 1998), pp. 329–332.