Speech Trends and Predictions, or Do We Need Text?

Patti Price

Speech Technology and Research Laboratory

SRI International

There has been remarkable progress in automatic speech recognition tec= hnology in the last ten years. Performance, however, remains significantl= y worse than human performance. Speech recognition systems do reasonably= well for careful speakers in relatively quiet environments when the cond= itions do not vary much from the training conditions. However, degradati= on can be severe with even minor levels of noise, with casual speech effe= cts, and with mismatches between training and testing conditions related = to dialect, vocabulary, and other characteristics. Nonetheless, speed an= d accuracy of the automatic systems have increased to the point where man= y useful applications are possible. The rapid growth in the number of co= mpanies devoted to speech recognition applications attests to this growth= in performance. This brief report explores the further potential for sp= eech technology.

Recent History of Performance Gains. Government benchmarks in = speech recognition have shown steady performance improvement. As perform= ance improves to below about 10%, the benchmark tests have expanded to be= more challenging. The community has moved from tasks focused on speaker= -dependent performance to speaker-independent (and adaptive) performance,= from small vocabularies to larger ones, from a sound-attenuated room to = telephone and broadcast news speech, from carefully read speech to sponta= neous speech, from simple tasks to more complex ones. While it is true t= hat computational power has increased and memory costs have decreased, it= is also true that researchers have consistently been able to make use of= the increased computational cycles and cheaper memory. Although the gov= ernment benchmarks do not assess speed, applications demand speed; the co= st of telephony applications depends largely on how much throughput can b= e handled on a server. The fact that the number and the size of speech c= ompanies and of speech groups within other companies are growing seems to= be evidence that the demands of speed and size, as well as of accuracy, = are starting to be met.

Humans Outperform Machines. Several studies have compared huma= n to machine performance in speech recognition, and in general these stud= ies suggest that humans are far superior to machines in accuracy. Many o= f these tests, however, use several listeners and allow multiple listenin= g passes to obtain "human" performance. One can wonder, then, = how humans could ever be wrong, since they are determining the "corr= ect" response. Further, it has been shown in recent experiments tha= t system performance can also be improved by combining the outputs of sev= eral different recognition systems. This is evidence that our systems are= not quite so highly correlated as has sometimes been claimed, and that t= here is still room to learn from each other. These details aside, it is = generally true that humans outperform our very best systems, particularly= when faced with casual speech, with dialects not well represented in the= training materials, and with relatively low noise levels.

Machines May Outperform Humans. Despite their generally superi= or performance in measures of

accuracy, humans have some disadvantages re= lative to machines: They want more in salary and benefits than many syst= ems require, they are not willing to work 24 hours a day, they can suffer= from inattention and boredom, and they are not as good at parallel proce= ssing. In many cases they are not as fast as machines. In addition to t= hese general differences, there are probably now some other cases in whic= h machines can outperform humans, though I know of no data yet to this ef= fect. For example, machines may be no more degraded in noise than are (o= ther) nonnative speakers, they are probably better than humans at speaker= identification based on small training sets, and they are probably bette= r than humans at recognizing nonspeech characteristics, such as channel e= ffects.

Current Speech Recognition Applications. The two main areas in= which speech is being commercialized are eyes/hands busy (over the telep= hone and/or in a car) and dictation. These are good first applications, = because they enable the user to interact with and to adapt to the capabil= ities of the machine. People are far more adaptable at present than are = our systems, and when recognition errors are seen, they adapt the way the= y are speaking to reduce these errors. We have seen a halving of error r= ates in the second compared to the first ten minutes of interaction, for = example. The reduction in error rate seems to arise from people using fe= wer out-of-vocabulary/domain terms, speaking more carefully, and speaking= more fluently.

The fact that people can adapt to the systems makes them more tractabl= e at present, relative to conditions in which a person cannot adapt.

Future Directions.

Spoken language is the medium used first and foremost for interact= ive human communication. It may be what distinguished us from the beasts= in allowing more information sharing. This information sharing was, how= ever, limited by life span and memory. The move from the oral stage to th= e written stage of communication was an information revolution, because i= t enabled communication at a distance in time and in space (if you could = remember where the book was and could find the information you wanted wit= hin it). The computer age brought another revolution in making online te= xt randomly accessible. It also brought new sources of information: for = example, radio, video, multimedia. Most of the currently commercialized= applications of speech recognition use speech as a means to access infor= mation (database query, command and control) or purely to input informati= on (dictation). However, I=92d like to ask here that we imagine that we = could endow speech with all the properties that we now enjoy in online te= xt. I am not yet ready to give up my book by the fireside, but as techno= logists we can learn something by asking: If we could give speech these a= ttributes, would we need text? Setting aside instances in which the audit= ory channel is simply not available, what makes us choose text over speec= h and can we imagine it otherwise? Imagine that all speech could be acces= sible as an information source. What types of speech would be of interes= t, how would it be used, and what technologies are needed to enable this = vision?

Some types of speech that might be of interest are voice mail messages= (particularly for call-in help lines), talks at technical, planning and = design meetings, radio and television broadcasts, and speeches. In fact, = if speech were as accessible as text, perhaps it would be used more. For= example, speech annotation

could be used in documentation of computer pr= ograms, of design decisions, or of video. Video annotation is of interes= t so that we can retrieve video segments even though there may be no acco= mpanying speech to transcribe, no speech that describes the video segment= of interest, or no video technology adequate to retrieve the item of int= erest. To enable access to these types of speech as an information source= , we will need further development in modeling spontaneous and casual spe= ech effects, speech in various kinds of noise, and speech from very heter= ogeneous sources. We can quickly scan text visually to find a section of = interest. Could we ask our technology to find a section of interest by a= sking: "When did X say Y?" Could we use speech technology to p= lay the section through quickly (blind people listen to synthesizers at v= ery fast rates). Often when we scan a text by eye, we find something of = interest that we didn=92t know we were looking for. Can we ask our techn= ology to find examples of things we previously found of interest, or thin= gs that are different or salient in a particular segment? Often we want t= o look through a document to get an idea of what it=92s about to know whe= ther we should read it in detail or recommend it to someone else. Can we= ask our technology to gist or to summarize the contents of a spoken docu= ment?

Text clearly differs from speech, and it is important to understand th= ose differences. When Sartre lost his sight, he stopped writing because w= riting was so visual for him. Can we develop technology that would have a= llowed Sartre to continue "writing"? One major difference is th= at speech unfolds linearly in time. It is harder to go back and reread. = Perhaps technology could make relistening easier. Text tends to differ = grammatically from speech. On the one hand it is simpler in that the rul= es seem to be more regularly applied. On the other hand, written languag= e tends to be more complex since writers can pack in more information bec= ause they know that readers can easily reread. This is part of what make= s presentations that are read rather than "spoken" so hard to u= nderstand: the information density tends to be greater (they may also be = more difficult to understand because the prosody may be infelicitous). Pe= rhaps technology can provide tools for spoken document production,= tools that enable the creation of coherent, easily understood audio docu= ments. Perhaps technology can provide tools that can transform one medium= (e.g., the spoken form) into another (e.g., the written form), including= an accounting for all the normal differences between the two styles: rem= oval of disfluencies, and repackaging of information. Perhaps technology= can translate a spoken document into a spoken document in another langua= ge, or translate a spoken document into the same language but with simple= r words and grammar so that it is more accessible to beginning readers or= nonnative speakers. In sum, let=92s not give up our books and text yet, = but let=92s ask why we need text now, and what technology could do to mak= e endow speech with those properties we appreciate in text.