One DODer's View of ARPA Spoken Language Directions

Calvin Olano

Department of Defense
Attn.: R523
Fort George G. Meade, MD 20755
colano@charm.isi.edu

1. Introduction

DOD support for ARPA/HLT speech research stems from the belief that large vocabulary continuous speech recognition and speech understanding will have advantages in portability, and in the variety of applications sustainable from a single, basic speech system. With some imagination one can envision applications in such remotely related areas as speaker identification and language identification.

Like many of you, I feel that some useful applications can be found for current systems, but many potential applications will fail the usefulness test. Those that pass will require considerable algorithmic tuning. So, how do we expand the range of applications, while simultaneously making life easier for the person who must develop applications? Of course, ARPA has been working toward this by introducing stress testing, unconstrained vocabulary, multi-lingual speech processing, contrastive testing, etc. But we can do more. I share a number of opinions with my colleagues in DOD about current, and potentially new ARPA research directions. Perhaps surfacing these perspectives will stimulate discussion and have some impact.

If we could agree on the goal of ARPA speech research, agreeing on the research directions might be easier. For me this goal is a simple motherhood statement. I would like to have a competent large vocabulary continuous speech recognition engine.

To achieve that competence we will need better basic performance that is robust to changes in problem and environment. Preferably the system would be capable of assisting in tracking changes in a problem over time. Once the initial system is trained, an adequate port to a new problem should be possible with limited training. Improved performance should be achievable over time by bootstrapping techniques that require minimum user interaction.

However, even if everyone accepted these goals there would still be sources of conflict. The first area of potential conflict I will consider is the relative emphasis of research versus technology transfer.

2. Tech Transfer vs Research

As we have long recognized, there are several ways to improve performance on a task. One is by improving the underlying science and algorithms. Another is taking advantage of natural structural constraints that are common to many tasks. But equally important is tailoring the system to use constraints and biases unique to the specific task. To clarify this view, consider an application like the ATIS speech understanding task: Finding features that provide a better performing talker independent system is an algorithmic improvement to the acoustic recognizer. Using knowledge of the prior discourse and likely new queries to guide the acoustic recognizer is a natural structural constraint. Designing a talker specific system because only a small set of talkers will use the system is tailoring based upon task specific biases.

I am deeply interested in task specific tailoring, and I don't care whether the information that guides the process comes from a speech or some totally unique task specific bias.

To advance speech research we work on general problems, like ATIS and WSJ, which must be

constructed with extraordinary caution to avoid bias. But, I love biases, as long as we know about them and can take advantage of them. And my experience is that most problems have biases that may be exploited once they are known. Perhaps some of our work should attack natural or real tasks with a no-holds-barred approach.

Does this mean that speech researchers should give up their careers and become system developers? -- Certainly not. But, at this point in time, throwing a speech algorithm over the fence to the system developer does not work. Some percentage of the speech knowledgeable workers will either have to become system developers or must establish a rapport with the system developer that is a top priority of their work. In the current state of development, only by tailoring the algorithm to the constraints of the problem will speech technology become cost effective.

It is my belief that working on more realistic problems, and being forced to think more about the technology transfer issues will produce systems are more capable of taking advantage of the specifics of a new task. Such a system would be architecturally different from a system designed to work as the general continuous speech transcription system or the general database interface tool. I believe that in addition we would further our understanding of our speech recognition systems and of human - computer speech communication.

3. Natural Constraints

We need to make money with our technology, but I have already expressed the belief that the number of applications we can expect success on today is limited. What are the limitations of current system performance? Can we take advantage of the natural structures in speech to improve performance?

Consider the speech tasks worked on under the ARPA HLT program. The Wall Street Journal transcription task can be useful in advancing large vocabulary speech recognition, but will viable applications of this technology follow soon? I'm skeptical. Considerable work has been done to improve underlying speech processing systems by using natural constraints such as statistical grammars. We need to do more. I do not know whether systems with little or no semantic or higher level knowledge can ever give acceptable performance on a real application like WSJ. Worse than that, I don't know how to measure the performance limits imposed by this handicap.

The alternative is not to live with the handicap, but to begin to introduce higher level knowledge, perhaps in a fragmentary manner. Indeed, some researchers are working to constrain lower level processes by using structure based on higher level considerations. Since I believe that this will prove to be very important, I would like to encourage this work. In addition, we should give thought to whether there are better architectures for applying these constraints.

4. Improving the Basic System

Since everyone is continually trying to improve their system, you might think that there would be little new to surface here. But this is not the case. There hasn't been enough time, enough money or enough manpower to do as thorough a job as we would like.

Improving the understanding of our algorithms is always a useful activity. I strongly suspect that many of the research elements present here do not understand the relative contributions and dependencies intrinsic to each of their system components as well as they believe they do. Discovering these dependencies is an evolutionary process. In most cases, many diagnostic tests that could be run to gain insight into the system have never been done. -- At a minimum, they have never been reported.

One aspect of system performance that is not often measured is consistency across talkers, channel environments, etc. Although we look at changes in word recognition performance, how often do we measure the consistency of our recognizers at the subword level?

We can learn more from our experiments and our data bases, if we are willing to make the effort. There are uncontrolled variables, the effect of which could be measured. For instance, we could use channel simulation to gauge the effect of channel differences. As another example, consider the WSJ task. There are great disparities of performance from talker to talker. If we reused our test data and restructured our test to reflect this "great divide", we would gain new insight. -- Are there other ideas for getting more milk from these data?

We should explore new testing paradigms. It would be useful to know how our systems differed from human performance. It should be possible to do psychophysical measurements on our systems and compare them to human psychophysics. As one simple experiment we could compare human and system performance using diagnostic rhyme, or nonsense syllable tests. This can be looked on as a contrastive test of human versus acoustic recognizer performance with higher level knowledge denied both humans and machines.

A more dramatic experimental change would be to run our systems with an "infinite" corpus of data. That is, we would devote a portion of our energy to testing our systems on a continuing, day to day basis on a realistic problem. We would get experience with how problems drift with time, and undoubtedly improve portability. This would also be a good scenario for having the speech systems perform certain levels of self diagnosis. The machine could tell us when discontinuities in the data occurred, and could be structured to assist in learning the necessary repairs. One advantage of this testing paradigm is that we would test our systems on orders of magnitude more data than we do in the normal static test mode, thereby obtaining more exposure to low probability events that could be saved for further study and system updating.

In the infinite data paradigm our view of training would change drastically. We would be blessed by having much more data available for training, and cursed by having less information about the data. From my point of view, it would be extremely beneficial to see the ingenious mechanisms that would evolve to cope with and take advantage of this situation.

5. Concluding Remark

I would like to hear a serious discussion of what we might do differently and what the benefits would be. A related issue that should be seriously addressed is what can be done to encourage more diversity of approach, more risk taking, and, consequently, more innovation.