Probabilistic Methods in Automatic Speech Recognition

A

Paul De Palma
Gonzaga University, USA

INTRODUCTION

Most research into Automatic Speech Recognition (ASR) assumes with Frederick Jellinek (1997), one of the pioneers in the field, that a "speech recognizer … can be thought of as a voice-actuated 'typewriter' in which a computer program carries out the transcription and the transcribed text appears on the word display" (p.1). Though this modest formulation rather understates what we humans do when we engage in conversation, perhaps the ultimate goal of ASR, we can use it as a working definition of contemporary systems. Still, Jellinek's simple description of *what* a speech recognizer does belies the complexity of *how* it does it. ASR is interdisciplinary by its very nature, spanning digital signal processing, acoustics, phonetics, information theory, probability and statistics, software engineering, and machine learning. It can appear quite forbidding to non-specialists. This is too bad, since the central insight of ASR—that speech is a probabilistic phenomenon—relies on a set of techniques developed by an English parson over two centuries ago. Filter this insight through a little digital signal processing, package it in modern software, run it on inexpensive hardware and you have Apple's Siri, Google's voice apps, and Nuance's Dragon Naturally Speaking. This is not to discount the complexity of ASR. Language is one of the distinguishing features of our species. To formalize even a small part of it requires substantial training, cleverness, and just plain persistence, persistence of the sort that required a full four decades for the first genuine breakthroughs in ASR to appear. Nevertheless, the basic contours of the field are accessible to non-specialists. Providing a readable account of ASR, while doing justice to the complexity of the field, is the goal of this article.

BACKGROUND

Work in automatic speech recognition has a long history. It began at about same time that researchers first developed compilers for the early high-level programming languages. Bell Labs, RCA Research, and MIT's Lincoln labs all used new ideas in acoustic phonetics to work on the recognition of single digits, syllables, and certain vowel sounds. Work continued through the 1960s in the United States, Japan, and the Soviet Union, using pattern-recognition techniques. This work received a big boost after the development of linear predictive coding in the 1970s, a technique to represent a compressed version of an acoustic signal. In all cases, however, the effort was to develop systems that could recognize single words. Two developments in the 1980s gave ASR its modern shape. The first was Defense Advanced Research Products Agency funding for research into large vocabulary continuous speech recognition (LVCSR). LVCSR systems have vocabularies in the 20,000 to 60,000 word range and process continuous speech from multiple speakers. The second was the adaptation of statistical techniques, most notably hidden Markov models, to the speech recognition problem.

Though current ASR falls far short of the bar established in the 1968 movie, *2001*, where a malevolent, but graciously conversational computer, takes control of a space station, converting acoustic waveforms to text is useful in a wide variety of contexts. The most obvious community to benefit from voice-augmented computers is the sight-impaired, for whom it might be argued that the evolution from textual input to point-and-click devices has made interacting with a computer somewhat more difficult. The rapid development of mobile computing devices is shifting the user

DOI: 10.4018/978-1-4666-5888-2.ch024

paradigm away from a person sitting at a desk, and possibly away from point-and-click itself. Speech is a natural replacement; but the movement from desktop to mobile computing does not exhaust the possibilities for a spoken interface. Anytime a computer user's eyes or hands are not available, as in situations where equipment or objects must be handled, the point-and-click model has clear limitations. From automobile mechanics, to industrial device operators, to medical technicians, to airplane pilots, to surgeons, any application that requires hands, eyes, and a computer is a candidate for speech recognition (and synthesis, of course).

The history of computing is the history of replacing relatively low-skilled occupations with software and software-mediated hardware. So, web-based computer interfaces have transformed the travel and consumer banking industries in the past decade, while American factories produce more with fewer workers because of productivity-enhancing technologies. On the other hand, even though advances in telecommunications have made the off-shoring of call centers to low-wage countries with an abundance of English speakers very attractive, Indian call center operators must still be paid. Dialog systems, which begin with automatic speech recognition, are obvious extensions both to web-based interfaces and to human-operated call centers. Conversational dialog continues to be an area of active research in the speech recognition community.

One area of ASR that has been quite successful is dictation, the text transcription of an extended monologue by a single speaker. The leading ASR in this area has been the software suite, *Dragon Naturally Speaking*. The drawback, however, is that such systems must be trained to a single speaker's voice.

One should not be surprised, given the history of funding for ASR research, that a potential beneficiary is the surveillance community. Just how the National Security Agency will process the data it is thought to have collected is rarely discussed. Though the workings of the NSA are classified, of course, it is easy to imagine that it employs voice recognition software to scan phone calls looking for words that indicate a threat to national security.

Before going into the details of ASR, it is important to point out that various task-related factors affect the accuracy of systems. For instance, ASR in a noisy environment is more difficult than ASR in a quiet room. The more constrained the vocabulary, the better the performance. And ASR performs better on read speech than on continuous, conversational speech. This should be no surprise. Foreign language learners the world over perform better or worse in exactly the same settings as ASR systems. The gold standard is large vocabulary continuous speech recognition. Solve this problem and more constrained problems are solved as well.

THE ARCHITECTURE OF LARGE VOCABULARY SPEECH RECOGNITION SYSTEMS

Speech recognition systems have a four-part architecture indicated by the ovals in Figure 1.

Modern ASR assumes that an acoustic waveform can be treated as a noisy, i.e., partially scrambled, version of a pre-existing string of words. If we could figure out how the input was scrambled, we could build a model that maps the waveform to the string. The central problem for ASR then becomes the answer to this question: "What is the most likely string of words of all word strings in the target language given some acoustic input?" Since digital signal processing technology lets us sample speech at regular intervals, we can treat these samples as a sequence of individual acoustic observations:

$$O = o_1, o_2, \ldots, o_t$$

In the same fashion, we can treat output as a string of linguistic symbols conditionally dependent upon the input. Words, syllables, units of sound called phones, phones with their left and right contexts called triphones would all suffice, though it is easiest to imagine a simple word string:

$$S = s_1, s_2, \ldots, s_n$$

In effect, among all candidate word strings we would like to find the one that is most probable given the acoustic observations. This is expressed in Equation 1:

$$hyp(S) = \frac{max}{S \in L} P(S|O) \quad (1)$$