Reasearch and Teaching Interests

John McDonough
Language Technology Institute,
Carnegie Mellon University
5411 Gates Hillman Complex,
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

January 1, 2012

My primary research interest is in statistical pattern recognition in general, and speech recognition in general; more specifically, I am especially interested in the emerging subfield of *distant speech recognition* (DSR), wherein it cannot be assumed that the microphone used for speech capture is directly adjacent to the speaker's mouth; rather, the microphone might be located several meters away from the mouth. This topic is of interest to me, because solving it requires the solution of several problems not encountered in conventional speech recognition, including dealing with the effects of noise, reverberation, and competing speech. The DSR problem is also necessarily interdisciplinary, and shares elements with statistical pattern recognition, digital signal processing, finite-state automata theory, and natural language processing.

My first exposure to DSR came while I worked at BBN Technologies in Cambridge, MA. From 1993 to 1997. My initial work involved implmementing algorithms for unsupervised speaker adaptation, whereby the characteristics of a given speaker's voice are learned, typically in an unsupervised fashion. I also developed techniques for applying speaker adaptation techniques during parameter estimation for hidden Markov models. In the Fall of 1997, I entered the graduate program in the Johns Hopkins University at the Center for Language and Speech Processing; my advisor was Fred Jelinek. At JHU, I continued to work primarily on speaker adaptation algorithms, for which I was awarded a Ph.D. degree in April, 2000.

From January, 2000 I took a job as a postdoctoral researcher at the University of Karlsruhe in Karlsruhe, Germany. In 2002 I attended the ICASSP conference in Orlando, FL and bought the book *Optimum Array Processing* by Harry Van Trees, which had only recently been published. Although well over one thousand pages in length, the word "microphone"—to my knowledge—does not appear even a single time in that work. Nonetheless, *Optimum Array Processing*, along with a compilation of research papers published earlier, inspired me to begin working on applying microphone arrays to speech recognition. To-

gether with a professor from the electrical engineering department, I set out to teach a course about exactly this topic, *Microphone Arrays: Gateway to Hands-Free Speech Recognition*. Admittedly, the first time we taught the course, of which I held the vast majority of the lectures, I was reduced to essentially repeating what was written in Prof. Van Tree's big book. Nonetheless, the students reacted positively, and several of them later came to work at our lab; their work eventually resulted in several research papers that we were able to publish in international journals and conference proceedings. After a few years of working in the field, I acquired a relatively good understanding of what portions of the traditional array processing literature are useful for acoustic array processing, and which not so much. I found this to be a very rewarding experience; because I found several good students through the process of holding a course, I didn't at all rue the time spent in preparing lectures, homework assignments, and mid-terms; rather, I viewed it as a prudent investment.

In 2004 I became involved in the CHIL, *Computers in the Human Interaction Loop*, project at the University of Karlsruhe. This project was ideal as a means of exercising the skill set I had been developing for the past couple of years, as it took as its goal the construction of "smart rooms" or "interactive spaces" that had the capacity to understand speech and spontaneously assist human users in interacting with other users. As part of our project work, we collected several substantial corpora of far-field speech data, which were then used for DSR experiments among other things. We used these corpora both within the project, but also eventually shared it with the US *National Institute of Standards and Technology* (NIST) for their use in a series of annual technology evaluations. NIST subsequently provided the data to other top research sites including LIMSI in Paris, France, IBM Watson Research Center in Yorktown Heights, NY, the University of Edinburgh in Edinburgh, UK, and Carnegie Mellon University in Pittsburgh, PA.

Some of the most promising techniques to have come out of our research involve the use of the non-Gaussian properties of human speech for beamforming, which by definition is the combination of all signals coming from a microphone array so as to focus on the desired speech of a speaker, while suppressing unwanted speech, noise and reverberation. When examined in either the time for frequency domain, it becomes readily apparent that speech is not Gaussian, but highly super-Gaussian; this implies that its *probability density function* (pdf) has a great deal of mass centered around the mean and in the tails of the distribution, but relatively little mass in the intermediate regions. Due to the central limit theorem, when speech is corrupted by noise, reverberation or the speech of a competing speaker, the resulting pdf begins to approach that of a Gaussian. My students and I discovered by that combining the signals from a microphone array so as to restore the statistical characteristics of the original, uncorrupted speech, a class of algorithms can be obtained that yields DSR performance that is superior to any conventional beamforming techniques.

More recently, we have become interested in beamforming algorithms for *spherical* microphone arrays. The latter are of interest because they provide a directional sensitivity pattern whose shape is not altered when the direction

of interest is changed. The theory behind spherical microphone arrays is quite interesting, in that mastering requires learning something about room acoustics; hence, we've recently been actively reading about this field. In support of a book chapter we are preparing, we plan to report the results of a set of experiments that directly compares DSR performance for a conventional, linear array, with that of a spherical array. We actually now have several publications going back half a dozen years comparing beamforming algorithms in terms of *word error rate* (WER), which at times has caused a few raised eyebrows; beamforming performance is typically measured in terms of *signal-to-noise ratio* (SNR). We have found, however, that SNR does not correlate well with WER; if the latter is what is wanted, it makes no sense to measure the former.

Several other supporting technologies may also be required to build a complete DSR system. For reasons of computational efficiency and speed of convergence, beamforming is typically applied in the *frequency* or *subband domain*; hence, digital filter banks are required for performing subband analysis and—subsequent to beamforming—resynthesis. As I learned after some time working the field, the type of filter bank best suited to adaptive filtering and beamforming is very different from the filter banks used for data coding and compression, in that the former cannot be based on the notion of *aliasing cancellation* as a the latter often are. Prior to the effective application of beamforming, it is necessary to either *know* or *robustly estimate* the position of the desired speaker. Speaker tracking is typically performed with some variant of a *Kalman* or *Bayesian filter*, often using *time delays of arrival* (TDOAs) as input observations; over the years, we've developed several such speaker tracking systems. Moreover, if a DSR system is intended to *interact* with a speaker as opposed to simply recognizing his or her speech, it must provide for a "barge-in" capability; this in turn implies that some sort of *acoustic echo cancellation* (AEC) must also be present to suppress the system's voice prompt in the captured speech. In recent work, we've compared AEC performance based on the conventional *normalized least mean squares* (NLMS) algorithm, with more sophisticated techniques based on a Kalman filter. We've found that the normal, covariance form of the Kalman filter works better than the NLMS algorithm, but that the information form of the Kalman filter works better still. The latter fact stems from the fact that a regularization term can readily be applied to the information filter in order to control the growth of the subband filter coffificients, and thereby obtain better robustness.

All of the above deals almost exclusively with the signal processing aspects of distant speech recognition; but the problem of actually finding a word sequence that best matches the acoustics must also be addressed. In considering this problem, I've found the most intellectually satisfying body of techniques to be those based on the theory of *weighted finite-state transducers* (WFSTs). This set of techniques was derived from the conventional techniques for manipulation of finite-state automata (FSAs), such as set intersection and power set construction. The conventional automata are generalized through the inclusion of a *weight* on each arc along with the usual input symbol, where the weight can be an output symbol, a real value representing a probability, or the

Cartesian product of both. Such structures are useful for speech recognition because all of the structures required for the latter can be represented as WFSTs, which can then be combined through weighted composition—the analog of set intersection—then optimized through weighted determinization—the analog of power set construction. To obtain further reductions in run-time, the weights can be pushed towards the start node, then the result can be processed with standard FSA minimization after encoding the input symbol, output symbol and weight on each arc as a single symbol.

My work during the CHIL project on acoustic array processing was in many ways a defining event in my career, because largely on the basis of that experience I was able to assemble enough material to co-author the book *Distant Speech Recognition* (DSR) with a former student; this book was published in April, 2009 by Wiley. DSR represents the culmination of everything I've learned about this topic after working in the field for nearly 20 years. At some point, I hope to publish a second edition of DSR to report on everything I've learned since it's original appearance, as well as what I've learned by using DSR as a text book at both Karlsruhe and Saarland University; at Saarland University, I've used DSR as the text for both an eponymous course, as well as second course entitled *Weighted Finite-State Transducers in Speech and Natural Language Processing.*

Let me summarize by saying that my interest in DSR is motivated also by the tremendous variety of applications that would be made possible by such technology; to name only a few:

- Speech-enabled video games;
- Co-operative robots;
- Car navigation and information systems.
- Smart houses and home entertainment systems.

Building DSR systems that are sufficiently robust and capable to work reliably in such applications is not a problem that will be solved in two or even five years. Nonetheless, these are applications of current interest in both academia and industry.