Embedded Speech Recognition Applications in Mobile Phones: Status, Trends, and Challenges

Jordan Cohen
SRI International
333 Ravenswood Road, Menlo park, CA. 94025
jordan.cohen@sri.com

ABSTRACT

Voice centric interfaces are widely available in modern mobile phones, including low-cost versions. The applications have evolved from speaker-dependent name dialing, which require user enrollment of frequently dialed names, to speaker-independent capabilities including continuous digit dialing, command and control of phone functions, and name dialing directly from the phone's contacts directory. Recently available advances include capabilities like voice-enabled SMS, e-mail, and even mobile search with voice. This evolution has been enabled by advances in speech recognition robustness, network capabilities, and increased computational power in small devices. Systems may now be used in hands-busy/eyes-busy conditions including speakerphone and bluetooth scenarios. In this paper, we will provide an overview of embedded speech recognition centric applications in mobile phones, specifically focusing on current status, industry trends, and challenges in customer acceptance. Although voice interfaces are natural and attractive in theory, a majority of users do not use the voice-enabled features available in their mobile phones. We will discuss some of the reasons for this user behavior and recommend actions to be taken.

Index Terms: Speech Recognition, Mobile, Applications

1. INTRODUCTION

Speech recognition technology has advanced tremendously over the last four decades, from ad-hoc algorithms to sophisticated solutions using hill-climbing parameter estimation and effective search strategies. While these algorithms advanced, mobile devices became ever more competent computing platforms, lagging desktop computers by only a few years. The combination of sophisticated algorithms and generous computing capabilities has not, however, put a speech recognition system in everyone's daily technical diet. We discuss briefly the advances in speech recognition, the computing scenario in portable devices (mostly cell phones), and the applications conundrum that has made these advances technical step children in the consumer driven economy.

2. EARLY INFRASTRUCTURE

In a 1948 epistle, Potter, Kopp and Green [1] demonstrated that speech sounds had characteristic patterns. They displayed "spectograms" of many sounds and syllables, made with the Sonograph, a mechanical spectral analysis and display device. It was clear from their book that the speech recognition problem was tractable.

Meanwhile, computing was coming of age. The IBM tabulators were morphing into computers, and by the 1960s, we had computers like the IBM 1620, a device with 4000 words of memory and a multiplier that operated by table look-up. Following Moore's Law, the 1970's saw the introduction of the Supercomputing PDP-10 and minicomputer the PDP-11 by Digital Equipment, and the signal processing age was upon us.

At the AT&T Laboratories, at Bolt Beranek and Neumann, and at the Speech Technology Laboratory in Santa Barbara, researchers studied the many nuances of linear predictive coding. This extraordinarily efficient computational algorithm allowed one to easily compute the resonances associated with a speech sound, ignoring the pitch, and confirmed the Potter Kopp and Green observations that sounds had "predictable" structure. Itakura and Saito noted that the spectrum derived from LPC (Linear Predictive Coding) highlighted the resonances of the vocal tract [2], and they defined the "Itakura Distance" which, under the right assumptions, expressed the distance between two spectra as an information measure. John Makhoul [3], in a very clear review paper, demonstrated that the solution of the LPC equations were a straightforward function of the autocorrelation of the speech signal. The computing load was compatible with the computers of the day, and LPC was born as the workhorse of speech research. Markel and Gray [4] produced a detailed monograph describing the relationship between LPC parameters and speech. Note that LPC parameterization is used in many speech coders, and forms the basis for many modern cell phone algorithms.

Given an LPC encoding of two speech signals, the signals could be compared by warping (adjusting the time of one to

line up with the other), and by accumulating information about the goodness of the match. The time alignment algorithm dynamic time warping (DTW) was an implementation of dynamic programming, promoted by both AT&T on the East Coast [5], and George White at Fairchild on the West Coast.

3. WORD SPOTTING AND SPEECH RECOGNITION

Once the speech signal could be efficiently characterized, the floodgates opened for speech applications including speech coding using LPC, word spotting, and speech recognition. During the 1970's, the government funded research and testing programs in "word spotting", where known words were to be identified in the speech of talkers unknown to the training algorithm, yielding time warping algorithms for matching acoustic utterances. In the same time frame, speech recognition support was provided by ARPA (the Advanced Research Projects Agency) under the SUR (Speech Understanding and Recognition) program, and for the first time systems were able to recognize complete sentences chosen from a finite state grammar.

We are ignoring here a branch of speech recognition based on neural networks which, although it has found use in small appliances and toys, has had small impact on mobile devices. See Gold and Morgan [6] for a more balanced review.

Another branch of speech recognition, the Hidden Markov Model(HMM), was developed at IBM and later HMM technology was embraced by others in the field. This technology was introduced to the community by IDA in a conference in 1982, but the IBM research organization had been exploring this space since 1970 in large vocabulary applications. Most modern embedded speech recognition applications use the HMM technology [7].

Hidden Markov Models allow phonetic or word rather than frame-by-frame modeling of speech They also are supported by very pretty convergence theorems[8], and an efficient training algorithm. Unlike DTW, the HMM recognition algorithms model the speech signal rather than the acoustic composite, and they tend to be more robust to background noise and distortion. In current applications, the cost associated with a mis-recognition are small enough so that sophisticated noise suppression techniques are not economically viable – it is often enough to train an HMM system with some noisy data.

4. THE RISE OF THE CELL PHONE

Meanwhile, the portable phone age was brewing. In 1973, Martin Cooper, of Motorola, demonstrated a 2.2 pound self contained cell phone. This technology was packaged as a car phone (later a bag phone), and the mobile age was born in the 1980's. Cellular infrastructure changed from operator-driven radio to dialed service using analog and then digital service during the 80's and 90's. In the 1980's, early Dynamic Time Warping (DTW)-based speech recognizers were developed for car phones Two examples are the car phone dialer produced by Interstate Electronics, and another by AT&T called the Victory Dialer.

By the mid 1990's the US was entering the digital cellular era. Cell phones dropped to a few ounces in weight, and computing increased from abacus-like to processors running at a few Megahertz with some tens of thousands of bytes of memory.

The author owned a Motorola StarTac, a very early flip phone. It had SMS messaging and an internal phone book that would hold 2880 names (more than a typical cell phone today). It weighed 3.1 ounces, and was nearly indestructible. The StarTac came in many versions, and its six-year lifespan covered the conversion from the analog AMPS analog cellular system to TDMA and CDMA digital service in the United States.

Shortly after the StarTac, the author owned an Ericsson T28, a small GSM phone with a two line display. But it had voice dialing – speaker dependent name dialing from a single example provided by the user. This phone was an example of DTW speech recognition using likely an LPC front end, and had many of the characteristics which have delighted and maddened users of this technology to this day. The speech recognizer was simple to use, but had to be trained carefully. It did not work well in noise. You had to remember exactly how you said each name in training (Rob and Robert would not match each other), and the phone only held 10 names in its voice inventory. This type of speech recognition application was also available to the industry in Qualcomm, Motorola, and Texas Instruments chipsets, and after 2000 was made available widely by Advanced Recognition Technologies. Siemens, Ericsson, AT&T, and other companies were working on and advancing these applications

The advantages of DTW recognition – the audio signal associated with a name could be arbitrary, thus allowing use by people with accents – were offset by the disadvantages – the models were large and phones only supported a few names, and they performed poorly in noise. An alternative technology from the Hidden Markov branch of speech recognition started to occur in the early 2000s. Standard LPC front ends were replaced by Mel Cepstra or Perceptual LPC. Phonetic models replaced word models, and performance was enhanced. (On the other hand, algorithms expected "standard" pronunciations, and these algorithms did not support heavy accents or unexpected

pronunciations.) The computing power available in phones increased almost in lockstep with Moore's Law, but lagged the power of the desktops by a few years.

By 1992, cell phones had differentiated processors for the "cell phone" functions and for the user interface. The UI processor, mostly patterned after a very successful low-power series of designs from the ARM Corporation, were running at tens of Megahertz, and were programmable (at least by the manufacturer). The cellular telephone functions were increasingly partitioned from the User Interface processors, and became powerful but not programmable cellular system elements. The hardware availability of the user interface processors led the way to tremendous advances in speech recognition in cell phones.

5. THE ADVENT OF SPEAKER INDEPENDENT CELL PHONE RECOGNITION

In 2002, the Samsung A500 was introduced, containing a speaker-independent digit recognition system. It would allow the user to push a button, say "digit dial", and then after a verbal cue by the phone, to say a string of digits representing a telephone number. It required no training, and was the first HMM model recognition system available in a mobile handset.

In the A500, each digit was represented by an HMM word model. Training was done with a moderately large corpus of recorded speech. The phone was met with very positive reviews.

Shortly thereafter, the Samsung A600 became available, and it contained not only the digit dialing application, but a phonetically-based name recognition system allowing the user to voice dial any name typed in his phone book, as well as several command-and-control functions. No training by the user was necessary. More phones with these applications followed, both from Samsung and LG.

Many similar applications appeared in the early 2000's, some on cell phones, and others on smart phones or connected PDA's which allowed third party applications. None met with substantial commercial success [9].

6. NEW APPLICATIONS

With the advent of the ARM-9 processor in phones, it has been possible to support even more sophisticated applications. The Samsung P-207, launched in August 2005, contained a very competent speaker-adapted large vocabulary recognition system which allowed users to dictate SMS messages and email. The training sequence for adaptation to the talker was a series of 124 words cued by the phone.

The underlying technology is based on a Phonetic speech recognition engine using a Markov model, modified to be very efficient in both computation and footprint. While the details are closely held, this capability demonstrates that the current hardware can support a multitude of speech recognition applications. Among the applications currently being developed are navigation systems, voice enabled mobile search, and continuous dictation for text creation.

Because modern cell phones have multiple connections to the network, and because voice channels have increasing fidelity, many speech services are available through the network as well as locally. Many carriers offer voice dialing and messaging services by voice; the technological challenges here are operational, but the underlying speech algorithms and techniques have much in common with embedded systems.

7. CONSUMER INDIFFERENCE

Given the increase in local computation, the advances in speech technology, and the widespread use of mobile devices, why isn't modern speech technology obvious everywhere? Why don't a majority of the 1 billion people who bought cell phones last year use speech to manage their communications Three reasons occur to the author: Marketing, Applications, and Interfaces.

Marketing: The original voice applications were speaker trained dialing systems which were difficult to train (especially for the "blinking-12" generation), and which offered sub-optimal performance even when trained. The technology was oversold, and no recent marketing campaign has been able to differentiate between the modern, competent speaker-independent technologies and the earlier implementations. In the cell phone industry, marketing is controlled by the carriers, and they have not decided that speech interfaces are critical for their financial success. Without a marketing campaign, consumers continue to be poisoned by earlier bad experiences, and cell phones are so loaded with applications that consumers can't find speech applications even if they are positively inclined.

Applications: What is the must-have voice application? Voice dialing is very useful, and has increasing value as government organizations legislate that cell phones must be hands-free in the mobile environment. Bluetooth technology coupled with voice dialing satisfies the no-touch requirements of these legal restrictions, but the author's informal poll of local cell phone users suggests that the great majority of users do not use any speech-enabled applications. Users don't "need" the applications that have been developed. (The problem of performance for users

with accents has been solved in the laboratory, but has not been available to users, so there is some user dissatisfaction even with these modern systems. The consumer solution awaits an enlightened provider or the widespread availability of data).

Interface: Building an intuitive speech interface is a difficult task. Consumers are unwilling to train systems, and even if they are willing, the training is often unsuccessful because of the lack of sophistication of the average consumer. The only applications with any chance of success are those which are intuitive, which work out-of-the-box, and which deliver value to the consumer. It is possible that voice search will be that killer application, but the jury is still out.

8. RECOMMENDATIONS

Modern mobile devices can support a full complement of speech recognition technologies, but consumers have been ignoring them. Given below are specific recommendations to render speech recognition a "must-have" feature in mobile phones.

1. Develop a killer application, and let the application drive the technology development. Mobile devices are not general purpose computers, but almost any subset of the technological offerings of the speech community can be implemented in support of a targeted application. While further research in speech recognition and language technology are warranted, entirely competent resources exist for a huge collection of applications. A killer application is one which is so compelling that the user, once having the application, is unwilling to do without it. Examples are Visicalc for early computers, or the web browser for the World Wide Web. Defining this class of applications for speech systems has remained elusive.

2. Pay attention to the user interface. A great application with a defective interface will be a commercial failure. Apple has demonstrated the power of the user interface with the tremendously successful iPhone. (Note, however that the iPhone has NO speech recognition!) We have found that a good user interface must be intuitive – the application must capture the natural actions of the user and respond appropriately. While that sounds easy, in practice it is very difficult.

3. Tell the customers what they have. Marketing is an essential part of modern technology, whether direct or indirect. Viral marketing has been successful in some scenarios, and product placement can complement direct marketing campaigns. However, without a strategic marketing campaign, even the best application will be lost in the noise of technology overload in modern devices.

9. REFERENCES

[1] R. Potter, G. Kopp, and H. Green, Visible Speech, Van Nostrand, New York, 1947

[2] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," Reports of 6th Int. Cong. Acoust., ed. by Y. Kohasi, Tokyo, C-5-5, C17–20, 1968.

[3] John Makhoul, "Linear prediction: a tutorial review," Proceedings of the IEEE, Vol. 63, No. 4, April 1975.

[4] J.D. Markel, A.H. Gray, and H. Wakita, Linear Prediction of Speech — Theory and Practice, SCRL Monograph No. 10, Speech Communications Research Laboratory, Santa Barbara, California, 1973.

[5] C. S. Myers and L. R. Rabiner. "A comparative study of several dynamic time-warping algorithms for connected word recognition", The Bell System Technical Journal, 60(7):1389-1409, September 1981.

[6] B. Gold and N. Morgan, Speech and Audio Signal Processing, John Wilen & Sons, Inc, New York, 2000.

[7] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, 77 (2), p. 257–286, February 1989.

[8] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164--171, 1970

[9] J. Cohen, "Is Embedded Speech Recognition A Disruptive Technology?", Information Quarterly, 3 (5), 2004.