OPEN MIND SPEECH RECOGNITION

Jean-Marc Valin

Université de Sherbrooke
2500 boulevard de l'Université
Sherbrooke, Québec
J1K 2R1 CANADA
valj01@gel.usherb.ca

David G. Stork[*]

Ricoh Silicon Valley
2882 Sand Hill Road #115
Menlo Park, CA 94025-7022 USA
stork@OpenMind.org

ABSTRACT

We describe speech research through the Open Mind Initiative, which provides a framework for large-scale collaborative efforts in building components of "intelligent" systems using the internet. Based on Open Source methodology, the Open Mind Initiative allowsdomain specialiststo contribute algorithms, tool developers to provide software infrastructure and tools, and non-specialist "e-citizens" to contribute training data and information to large databases. An important challengeis to make iteasy and rewarding for e-citizens to provide such information. We describe the current status of such speech research, and several challenges and opportunities associated with the Open Mind Initiative.

1. INTRODUCTION

The development of classifiers and "intelligent" machines — ones that can understand speech, summarize stories, engage in conversation, etc. — relies on both theory [6] and data, and there has been incrementalimprovementin anumber of areas. Most subdisciplines in speech analysis and recognition require large corpora for progress; many would profit from an open framework for experimentation and collaboration. We discuss a new methodology — The Open Mind Initiative — to this end.

The paper is organized as follows: In Sect. 2 we stress the need for large corpora and an open framework for systems engineering and integration forfurtherprogress in several problems in speech processing and recognition. We then discuss the Open Mind Initiative, in particular its three components — domain experts, tool developers and non-specialist "e-citizens" — and briefly compare and contrast it with traditional Open Source. Then, in Sect. 3 we present currentworkon developing speech recognition systemswhich could beused withthe Open Sourceoperatingsystem *Linux*. Section 4 mentions some unsolved problems, research directions and conclusions.

2. THE OPEN MIND INITIATIVE

In very broad terms, recent work in many areas of pattern recognition and artificial intelligence has relied more and more upon fairly general models, such as powerful statistical ones, trained with a great deal of data. The fundamentaltheoreticalunderpinnings ofdomain-independentpattern recognition— maximum-likelihoodand Bayesiantechniques, function estimation, and so on — are highly developed and rigorous. While there willcontinueto beeffortandprogress, the foundations as currently understood are sufficient for developing successful pattern classifiers in many domains. The adequacy of even very simple models is illustrated in optical character recognition, where recognizers based on simple models (decision trees, neural networks, ...) trained with millions ofcharacters outperform recognizersbased on sophisticated models trained with less data [7]. This need for large training sets is a lesson that recurs in a number of domains, from acoustic speech recognition [9], speechreading [13], natural language processing [5], speech production [15], and others. For many areas where we may not yet have adequate models, we nevertheless know how to broaden and improve classes of models — to include more degreesoffreedom to accountforsources ofvariation, to set parameters, and so on — given enough data. In summary, then, it appears that in many interesting domains, particularly speech, large data sets are necessary.

The appreciation of the need for large knowledge bases and training data has led to the construction of publicly available databases. The National Institutes of Standards and Technology (NIST), the Linguistics Data Consortium (LDC), and others have compiled large databases of training data related to speech, language, documents and other domains. A representative example is that of the Macrophone project, compiled by Texas Instruments, a collection of roughly 200,000 utterances of free telephone speech from non-specialists, constrained by topic [2]. While these and other public databases have been vital to continued improvements in recognizers, some of the best systems are

[*] Presentingauthor

trained with yet additional data, usually proprietary, as in the case of a leading optical character recognition product [3].

Another key component in building such systems involves software tools. There are many commercial tools for developing speech and natural language systems which allow developers to explore model parameters easily, specify grammars, lexicons, and so on. Some of these tools are provided free of charge by companies in order to promote the sale and use of their hardware, such as speech chips [14, 8]. An important lesson here is that non-specialist developers can create useful systems, given sufficiently good tools.

It was in appreciation of these needs, and the success of the Open Source model of software development, that let to the proposal of the Open Mind Initiative [12, 11]. It is perhaps simplest to understand the structure of the Open Mind Initiative in terms of its three component functions — provided by domain experts, infrastructure and tooldevelopers, and e-citizens. Domain experts contribute libraries of fundamental algorithms; tool developers contribute and refine the enabling software; e-citizens contribute information and training data. All this is possible given the infrastructure of the internet and World Wide Web which lowers the effort and cost of creating large databases.

ThestructureofatypicalOpenMind developmentproject can be illustrated with isolated handwritten character recognition. In such a project, sampled handwritten characters are presented on web browsers of non-specialists who label (classify) them; the labels returned to the Open Mind host machine are used for training classifiers. The non-specialists are provided incentives and rewarded for contributing such pattern labels [11]. Machinelearning and pattern recognition algorithms are then used to train the classifier using the contributed labels. Domain experts can easily test different recognition algorithms and propose improvements, all in an open framework. The final software is downloadable and freely available to all.

2.1. Domain experts

Experts in a specific domain such as acoustic speech recognition contribute documented libraries of fundamental algorithms, and possibly representative training sets, freely available to all. Much of such work has already been published in refereed journals, and the Open Mind approach extends the trend in academic publishing in which algorithms and data are published in electronic form on the web.

2.2. Tool developers

Key components to the Initiative are provided by the infrastructure or tool developers. The challenge is to make it easy for e-citizens to contribute data, and here new forms of infrastructure will have to be developed. An unusual form of interface could be provided by games. In that case, tool developers would develop games in which the players's responses provided the feedback for training data. In such a relatively unstructured massive collaborative software project many technical problems must be considered — everything from low-quality data to outright hostile attacks. A number of simple heuristics in data "truthing" could reduce the possibility of poor data. For instance, any query from the Open Mind system could be presented to three independent, randomly chosen e-citizen contributors, and their replies accepted only if all three agree. Likewise, there are domain-dependent algorithms for automatically identifying "outliers" — responses that differ drastically from the current consensus, which could be automatically brought to the attention of a domain expert or moderator for review. There are many techniques from experimental psychology for insuring the quality of the data too, such as the insertion of a "catch trial" which has only one plausible answer; an incorrect answer on such a catch trial belies an unreliable contributor and invalidates his or her recent submissions. Finally, the software infrastructure should allow e-citizen contributors to identify themselves, for recognition and reward (see below).

2.3. E-citizens

The biggestdifference between traditional Open Source and the Open Mind Initiative is the need for data provided by e-citizens. E-citizens are non-experts — that is, they need neither programming skills nor academic knowledge of the problem domain. In essence, anyone with access to a web browser can be considered an e-citizen. One approach in the speech recognition domain is to play unlabeled sound files to e-citizens over the internet; each e-citizen would then classify the utterance, and indicate the label by clicking one of several graphic buttons on his or her browser. Even if a tiny percentage of people with web access provide a small amount of information, very large corpora could be created in this way. Furthermore, training could be efficient since the Open Mind system would present to e-citizens only ambiguouspatterns (i.e., mostinformative), atechniqueknown as learning with queries. Such learning often provides a distinct advantage over learning based on traditional i.i.d. sampling [1].

3. OPEN MIND SPEECH RESEARCH AND DEVELOPMENT

Although our speech project is in its infancy, we have identified some of its key initial components and began writing code. By the very nature of open development, these can be expected to change and improve as the project progresses.

Open Source	Open Mind
no e-citizens	e-citizens crucial
expert knowledge (e.g., data formats)	informal knowledge (e.g., phoneme identities)
machinelearning irrelevant	machine learning essential
web optional	web essential
most work is directly on the final software	most work is *not* on the final software
hacker/programmer culture ($\approx 10^5$ contributors)	e-citizen/business culture ($\approx 10^8$ contributors)
one or a small number of experts contribute a given part of the software (e.g., a *Linux* device driver)	many e-citizens contribute data that is used in a single function (e.g., digit recognition)

Table 1: Comparison of Open Source and Open Mind approaches.

Because of our location and personal interests, we expect that applications in both English and French will receive the earliest attention. The whole system should be portable to other languages, including "unicode languages" and Asian languages.

3.1. Recognition algorithms

The fundamental recognition algorithms are based on Hidden Markov Models and traditional grammars, as publicly known and available in the speech and pattern recognition literature [4, 9, 6]. Mostofour code is being written in C++; the first target operating system is *Linux*, though we want to make it easy to port our code to related operating systems such as *UNIX*.

3.2. Training data and tools

A key to developing automatic to speech recognition systems —anda centralconsideration in thedevelopmentofall Open Mind projects — is both the quality and the quantity of training data. It is important that data be collected for a wide range of applications. There is a need for simple command words ("yes," "no," "OK," "open," "start," ...), digits and digit strings, as well as more complex sentences for large vocabulary continuous speech recognition. This range demands that our data-collection tools be general purpose. During data collection, orthographic transcriptions can be entered by the speaker, reviewed by automatic "truthing" algorithms and if necessary by other e-citizens or domain experts. The audio data will be recorded at 16 bits/sample and 16 kHz sampling rate (PCM) to allow good quality. The compression (if any) must be lossless, based on publicly available algorithms.

Of particular importance to the Open Mind approach are tools for developing methods for collecting data from e-citizen contributors. Tool developers should be able to write and modify applications for playing unlabeled speech sounds to e-citizens and to collect their responses. We can envisionnovelgameinterfacestofacilitatecontributionsfrom e-citizens, for instance where the player's progress through a maze depends upon his or her labeling pre-segmented utterances played through the player's web browser. It will also be helpful to have the speaker enter additional information about the speaker (gender, age, dialect, native/non-native) and if applicable the recording method (microphone type, distance, noise level). The software tools should enables domain experts to manipulate and navigate the Open Mind speech database and extract data via queries that involve such supporting information. In this way, limited domain or age- or gender-specific recognizers can be developed.

3.3. Infrastructure for domain expert experimentation

Training datais important, butthere isalsoa need for agood testing database. While the collected data can be used for training or testing, it is important to organize the database so that data from a particular speaker is not used in both the training set and the test set. Along with each test set in the database, there could be a place for domain experts to note the performance gain/loss obtained be using different algorithms, so the information can be used by others.

4. CONCLUSIONS

There is of course a great deal yet to be done on speech and language processing through Open Mind. In addition to the numerous tasks outlined here, there is a need for new algorithms for outlier detection and data "truthing." Furthermore, in learning with queries, it is often desirable to *generate* new patterns for the e-citizen (serving as an "oracle") to label. Such patterns should be informative, i.e., such that the classifier classifies them with low confidence. This may require algorithms for generating speech utterances "between" two that already lie in the database, for instance an artificial utterance between a particular `/ba/` and a particular `/da/`. Some of these new algorithms may be of sufficiently general nature that they can be applied to projects in other Open Mind domains, such as optical character recognition.

The conjunction of several forces — the manifest need for"intelligent"software, thedemonstrated successofOpen Source methodology, the large body ofpowerfulmodelsand training algorithms, the infrastructure of the web and the very large and growing number of e-citizens — leads us to believe that the Open Mind Initiative provides an powerful framework for developing important and useful software. A particularly valuable aspect of Open Mind is that it facilitates integration ofseparate projects, for instance real-world ontologies, grammatical constraints, and topic identification with acoustic speech recognition. Open Mind seems to fulfill the need, broadly recognized within the artificial intelligence, speech recognition and related communities, for incorporating structure and constraints from such a wide range of functional and conceptual levels [10].

5. REFERENCES

[1] Dana C. Angluin. Learning with queries. In Eric B. Baum, editor, *ComputationalLearning andCognition* , pages 1–28, Philadelphia, PA, 1993. SIAM.

[2] Jared Bernstein, Kelsey Taussig, and Jack Godfrey. Macrophone: An American English telephone speech corpus for the Polyphone project. In *Proceedings of the International Conference on Automatic Speech and Signal Processing (ICASSP94)*, volume I, pages 81–84, Adelaide, Austrailia, 1994.

[3] Mindy Bokser, 1999. Personal communication (Caere Corporation).

[4] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, 1993.

[5] Walter Daelemans, Antal van den Bosch, Jakub Zavrel, Jorn Veenstra, Sabine Buchholz, and Bertjan Busser. Rapid development of NLP modules with memory-based learning. In Roberto Basili and Maria Theresa Pazienza, editors, *ECML98 TANLPS Workshop Notes*, pages 1–17, Technische Universität Chemnitz, 1998.

[6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, NY, second edition, 2000.

[7] Tin Kam Ho and Henry S. Baird. Large-scale simulation studies in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(10):1067–1079, 1997.

[8] Jerry R. Hobbs, Douglas Appelt, John Bear, and David Israel. FASTUS: A system for extracting information from text. In *Proceedings of the ARPA Human Language Technology Workshop '93*, pages 133–137, Princeton, NJ, 1994. Distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.

[9] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

[10] David G. Stork, 1999. Further information is available from `www.OpenMind.org`.

[11] David G. Stork. Document and character research in the Open Mind Initiative. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR99)*, Bangalore, India, 1999.

[12] David G. Stork. The Open Mind Initiative. *IEEE Intelligent Systems & their applications*, 14(3):19–20, 1999.

[13] David G. Stork and Marcus E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems, and Applications*. NATO Advanced Studies Institute. Springer, New York, NY, 1996.

[14] Stephen Sutton, Ron A. Cole, Jacques de Villiers, Johan Schalkwyk, Pieter Vermeulen, Michael Macon, Yonghon Yan, Ed Kaiser, Brian Rundle, Kal Shobaki, Peter Hosom, Alex Kain, Johan Wouters, Dominic Massaro, and Michael Cohen. Universal speech tools: The CSLU Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP98)*, pages 3211–3224, Sydney, Australia, November 1998.

[15] Antal van den Bosch and Walter Daelemans. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings ofthe Sixth Conference oftheEuropean Chapter of the ACL*, pages 45–53. ACL, 1993.