The Speech Recognition Virtual Kitchen

Florian Metze[1]*, Eric Fosler-Lussier*[2]*, and Rebecca Bates*[3]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA; U.S.A.
[2]Dept. of Computer Science and Engineering, The Ohio State University, Columbus, OH; U.S.A.
[3]Computer Science Department, Minnesota State University, Mankato, MN; U.S.A.
fmetze@cs.cmu.edu, fosler@cse.ohio-state.edu, bates@mnsu.edu

Abstract

This paper describes the "Speech Recognition Virtual Kitchen" environment which has the goals to promote community sharing of research techniques, foster innovative experimentation, and provide solid reference systems as a tool for education, research, and evaluation with a focus on, but not restricted to, speech and language research. The core of the research infrastructure is the use of Virtual Machines (VMs) that provide a consistent environment for experimentation. We liken the virtual machines to a "kitchen" because they provide the infrastructure into which one can install "appliances" (e.g., speech recognition tool-kits), "recipes" (scripts for creating state-of-the art systems), and "ingredients" (language data). A web-based community platform complements the VMs, to allow physically disconnected users to jointly explore VMs, learn from each other, and collaborate in research. In this demo, we present initial VMs that were mostly used for teaching classes at Carnegie Mellon and Ohio State University, and solicit feedback for an initial "hub"-style web-site.

Index Terms: speech recognition, virtualization, educational tools, research infrastructure

1. Introduction

Building and maintaining a state-of-the-art Automatic Speech Recognition (ASR) system has moved beyond the ability of a single developer; it is difficult for all but the largest of university laboratories to maintain an end-to-end system, and adapt it to new languages, tasks, or conditions as required. Speech recognizers incorporate knowledge from linguistics, phonetics, acoustics, signal processing, statistical modeling, graph theory, and artificial intelligence. Expecting students to become experts in all of these areas, before attempting to work on speech recognition systems, is unrealistic. This poses a high bar for developing new research groups, and makes it difficult for academic institutions without active ASR researchers to integrate ASR projects into their educational curricula or field research projects which include ASR, such as interactive dialog systems, speech-to-speech translation, human robot communication, multimedia analysis, etc.

What has been missing is a way for academic institutions (and industry) to leverage community resources in order to branch off new research from fully functional end-to-end systems, rather than a collection of individually downloaded tools and data. Even if a well-documented open-source ASR toolkit is being used, a "black box" approach usually results in poor performance. Some aspect of the system (audio data, text data, numerical libraries, etc.) could not be replicated exactly, or wasn't adapted to the requirements of a specific experiment — constituting a significant barrier for use by non-experts, and making it harder for experts to replicate each others' work.

This effort attempts to extend the model of lab-internal knowledge transfer to a community-wide effort through the use of Virtual Machines. We present a way to share entire "recipes", demonstrating how to build different kinds of systems, distributed ready-to-run, together with data, log-files, results, etc – everything that is expected from a baseline, in a working environment, and with links to other users that work on exactly the same task worldwide. Students and researchers can then modify recipes step by step, observing the effect of changes, or run the recipe on a different language, a different type of channel, or simply to perform a given test, in order to create a speech recognition system that is flexible and has good performance, and can be integrated in other, bigger projects.

2. The Speech Recognition Virtual Kitchen

The "Speech Recognition Virtual Kitchen" first serves as a repository for Virtual Machines, which will typically be based on redistributable operating systems such as the Ubuntu Linux derivative, to provide a common infrastructure the use of tool-kits and data. A user would download a VM from the "Kitchen Server" onto his "Host PC" as shown in Step ① in Figure 1, and run it. Initially, the VM may be generic and "bare" to reduce the size of the initial download, but it can be customized by installing additional software and/ or data (②) from either the "Kitchen Server" or third party servers. When installing from the kitchen repository, the user would install additional software or tools supporting a certain set of experiments, a tutorial, reference log-files, etc. Third-party servers will typically provide tool-kits, data, or simply additional packages and updates for the base OS. A package manager can be used, but is not required, making it easy to provide multiple different instances of "kitchen" VMs to the user – one kitchen will be an English LVCSR system trained using Kaldi, one will be a speech analysis workbench, one will be a language-independent keyword spotting system [1], etc. Because all VMs originated from the same configuration, there should be no compatibility issues – once an installation script has been found to work on one VM, it will on all VMs. The user can now reproduce experiments on his VM by changing the provided scripts or writing new code, and compare the results to the provided log-files and baseline results. By mounting local filesystems or connecting other local resources such as a microphone to the VM, the user can also access data on the host PC, which means that the "kitchen" experiments are not restricted to just supplied data. The kitchen however provides a locally available reference setup, against which any new installation (tools, data, etc.) can be verified, which has proven extremely useful in our own experience. The third

function of the "Virtual Kitchen" then is a "community" board (e.g., a Wiki, potentially with messaging and "high-score" functions, ③), which will allow users who downloaded similar VMs from the "kitchen" to connect with each other, and discuss their research, effectively porting a lab-based model of knowledge transmission to a global scale. We believe this configuration will be more effective than existing discussion boards, because users' installations will be much more similar to each other, so that no time is lost establishing things like "did you compile with this flag", "what happens when you run this baseline", etc.

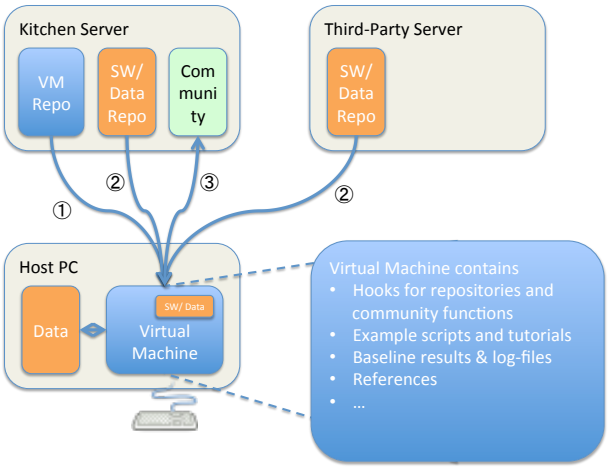

Figure 1: Architecture diagram of the "Kitchen" system.

Once configured, each virtual machine (VM) will be set up with instructions (typically a collection of scripts) for creating a speech recognition system that is toolkit and data source specific, and will contain the resulting system for reference. If separate licensing is required for specific toolkits or data, the VM can be configured to provide wrappers that make obtaining and installing the resources as easy as possible for the end user, while respecting the IP restrictions and distribution mechanism chosen by the original provider. If permitted by the license, data and software can be pre-installed on the end users' virtual machine fully automatically (by download from the "kitchen" server, or third-party servers), so that no extra steps are necessary before experimentation.

Under this paradigm, researchers and students can download virtual machines, easily reproduce an experiment, change a component (for example, the audio training and testing data) and re-run the experiment, observing the changes at every step, as is good scientific practice.

3. Show & Tell at INTERSPEECH 2013

At INTERSPEECH 2013, we will first demonstrate an initial prototype of the "Kitchen Server" web-site, and solicit community feedback on the interface, the desired functions, or other desiderate. We will also demonstrate a number of VMs that we have developed in the last year, which showcase the idea. At present, several VMs are available that have been developed during a succession of classes and labs at Carnegie Mellon and Ohio State University, and which can be shared with the community. For one VM (already showcased in [2]), students in the first year implemented a basic speech recognition system, and an interface to Second Life. In the second semester, students improved the dialog capabilities of the system, and added a face detector using the computer's web-cam to avoid the need for "push-to-talk". In the third semester, students added emotion recognition and performed experiments on a POMDP – all this without the need for a dedicated machine to host the environment, simply by passing the VM on from student to student. A similar approach has been followed at Ohio State, where exercises in building speech recognition systems were developed around the OpenFST framework [3] as well as the HTK toolkit [4]. We will demonstrate how VM technology enables easy distribution of educational materials, including the systems discussed above and other VMs that contain other speech recognition experiments and tool-kits, such as the open-source Kaldi [5]; we will also discuss our experience in developing and using them. Additional development will happen during the summer months, so that additional VMs, even some that have been provided by partners in the community, will likely be demonstrated.

4. Outlook

The "kitchen" has several broader impacts, which we plan to develop further over the next two to three years with the help of an upcoming NSF CRI (Community Research Infrastructure) grant. We seek community input to tackle some of the issues that have previously hampered efforts in cross-community sharing: intellectual property issues often stand in the way of widespread sharing, since different tool-kits and data may need to be licensed and thus are difficult to aggregate into one standalone package. We will define methods of distribution that preserve intellectual property rights, while a consistent environment will allow end users to install open or closed-sourced systems and data with consistent results. The Show & Tell at INTERSPEECH will provide an excellent forum for this, as did a similar event at INTERSPEECH 2012 [2].

We will also collect ideas for other scenarios, in which this infrastructure will be useful, beyond the ones discussed here. We believe that this infrastructure may be usable by fields other than core ASR that are data intensive (synthesis, dialog systems, NLP, computer vision, data mining), and that this may serve as an excellent example for future innovations. Incubating ASR in other fields by providing an easy-to-use, non-trivial research environment will boost the relevance of speech and language technologies in many other fields. The idea also has the potential to change the way that comparative results are reported: rather than just reporting a single number (word error rate) in comparing systems, researchers will be able to more easily compare the products of two recognition systems.

5. Acknowledgments

This material is based upon work supported by the National Science Foundation under grant CNS-1205589, "CI-P: The Speech Recognition Virtual Kitchen", and a follow-up full CRI grant.

6. References

[1] F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, and N. Rajput, "The spoken web search task," in *Proc. MediaEval Workshop*, Pisa; Italy, Oct. 2012, http://www.multimediaeval.org/mediaeval2012/sws2012/.

[2] F. Metze and E. Fosler-Lussier, "The speech recognition virtual kitchen: An initial prototype," in *Proc. INTERSPEECH*. Portland, OR: ISCA, Sep. 2012.

[3] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "Openfst: A general and efficient weighted finite-state transducer

library," *Proceedings of the Twelfth International Conference on Implementation and Application of Automata, (CIAA 2007), Lecture Notes in Computer Science*, vol. 4783, pp. 11–23, 2007.

[4] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Publishing Department, 2002. [Online]. Available: http://htk.eng.cam.ac.uk

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," in *Proc. ASRU*. Big Island, HI; USA: IEEE, 12 2011.