The Speech Recognition Virtual Kitchen: An Initial Prototype

Florian Metze\textsuperscript{1} and Eric Fosler-Lussier\textsuperscript{2}

\textsuperscript{1}Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA; U.S.A.
\textsuperscript{2}Dept. of Computer Science and Engineering, The Ohio State University, Columbus, OH; U.S.A.

fmetze@cs.cmu.edu, fosler@cse.ohio-state.edu

Abstract
This paper describes a “kitchen” environment 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 toolkits), “recipes” (scripts for creating state-of-the-art systems), and “ingredients” (language data). In this demo, we show an initial proposal for a community framework based on VM technology used for teaching classes at CMU and OSU.

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

1. Introduction
Developing and educating a next-generation workforce is becoming an increasingly challenging task, as building and maintaining a state-of-the-art speech recognition 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. 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 bar for developing new research groups, and also makes it difficult for academic institutions without active ASR researchers to integrate ASR projects into the educational curricula, or research projects where ASR is one component of a larger system.

What has been missing is a way for academic institutions (and industry) to leverage community resources in maintaining the state-of-the-art, and in ensuring that evaluations and comparisons are performed on meaningful and relevant tasks, rather than on a plethora of different datasets, tasks, and conditions. While open and closed-source ASR toolkits have become available to create quasi-replicable systems, treating a speech recognition system as a “black box” usually results in poor performance, as some aspect of the system configuration has not been 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 other’s work.

This effort attempts to extend the model of lab-internal knowledge transfer to a community-wide effort through the use of an innovative research infrastructure. We present an infrastructure to facilitate cross-site knowledge sharing: it is essentially a way to share “recipes”, demonstrating how to build different kinds of systems, ready-to-run and distributed together with data, log-files, results, etc — everything that is expected from a baseline, and a working environment. Researchers can then modify recipes to run 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.

The demonstration system we describe here serves two purposes: first, it demonstrates how virtual machine technology can be effectively used in the classroom by allowing rapid deployment of resources and consistent environments for coursework. The second purpose of the demo is to engage the community in thinking about a wider-scale community effort to build a common research infrastructure based on VM technology.

2. A Virtual Kitchen with “Recipes”
When aspiring chefs enter their training, it is clear that they are not developing their skills in a vacuum. Culinary academies provide a certain amount of infrastructure — that is, a kitchen — in which learning and experimentation can thrive. One can choose to place different kinds of appliances within the kitchen, and can cook with different kinds of ingredients. Recipes, and the skills to use them, are imparted by instructors to initiates. As chefs learn and grow, they take the basic techniques they have learned, and (hopefully) can create new and inspiring dishes by combining ingredients in different ways.

In the same way, training in corpus-based language technology depends on a set of appliances (in the case of ASR, speech recognition toolkits) and ingredients (speech data, transcripts, dictionaries, etc.), and cookbooks containing the recipes. However, established research labs will tend to provide training on one or two appliances, and smaller research labs may try to use an appliance but lack the cookbooks. What is needed in the field is a virtual kitchen that can provide the infrastructure needed for students to try out different techniques. This can lower the bar for entry by smaller research groups, and can enable non-speech labs to use speech technology in their own research beyond a “black-box” approach.

The “Speech Recognition Virtual Kitchen” serves as a repository for virtual machines (based on redistributable operating systems such as the Ubuntu Linux derivative) to facilitate the use of toolkits and data. 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. The end user is responsible for receiving a distribution of the appropriate ASR toolkit for the VM from the toolkit provider (by downloading or licensing the software), as well as the speech data. If permitted by the license, data and software can be pre-installed on the end users’ virtual machine fully automatically.
3. Show & Tell at INTERSPEECH 2012

At INTERSPEECH 2012, we will demonstrate a number of VMs that we have developed in the last year, which showcase the idea. Figure 1 shows a screenshot of a VM, which has been developed by students in several, subsequent classes and labs at Carnegie Mellon University: in the first semester, students 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 [1] as well as the HTK toolkit [2]. 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 toolkits, such as the open-source Kaldi [3]; we will also discuss our experience in developing and using them. We also plan to engage researchers and educators in higher education that typically do not participate in ASR research in a workshop at INTERSPEECH 2012 on this topic.

4. Outlook

The “kitchen” has several broader impacts, which we plan to develop further over the following months. 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 toolkits 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.

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 useable 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. Acknowledgements

This material is based upon work supported by the National Science Foundation under a CRI-P grant.

6. References