How to get started with speech-to-text?

Asked 16 years, 4 months ago Modified 10 years ago Viewed 4k times

10

I'm really interested in speech-to-text algorithms, but I'm not sure where to start studying up on them. A bunch of searching around led me to this, but it's from 1996 and I'm fairly certain that there have been improvements since then.

Does anyone who has any experience with this sort of stuff have any recommendations for reading / source code to examine? Or just general advice on what I should be trying to learn about if I want to get into the world of writing speech recognition programs (sometimes it's hard to know what to search for if you don't have much knowledge about the domain).

Edit: I'd like to do something cross-platform, but for the moment I'd be targeting linux.

Edit 2: Thanks csmba for the well-thought out reply. At this point in time, I'm mainly interested in being able to create applications that allow automation, or execution of different commands through voice. So, a limited amount of recognizable commands being able to be strung together. An example would be a music player that took commands like "Play the album Hello Everything by

Squarepusher", or an application launcher that allowed the user to create voice-shortcuts to launch specific apps.

I realize that it's a pretty giant problem, and that I have nowhere near the level of knowledge required right now to tackle implementing an entire recognition engine, although the techniques involved with doing so fascinate me, and it is something I'd like to work myself up to doing. In all likelihood, I'll probably end up picking up a book or two on the subject and studying up / playing with "simple" implementations in my free time.

language-agnostic **speech-recognition**

Share

Improve this question

Follow

edited Aug 18, 2008 at 17:12

asked Aug 18, 2008 at 16:05

jdd

4,336 ● 1 ● 28 ● 28

6 Answers

Sorted by: Highest score (default)

8

This is a HUGE questions, I wouldn't know how to begin... So let me just try giving you the right "terms" so you can refine your quest:

First, understand that Speech Recognition is a diverse and complicated subject, and it has many different applications. People tend to map this domain to the first thing that comes to their head (usually, that would be computers understanding what you are saying like in IVR systems). So first lets distinguise the concept into the main categories:

Human-to-Machine: Applications that deal with understanding what a human is saying, but the human knows he is talking to a machine and the *grammar* is very limited. Examples are

- Computer automation
- Specialized: Pilots automating some controls for example (noise a huge problem)
- IVR (Interactive Voice Response) systems like Google-411 or when you call the bank and the computer on the other side says "say 'service' to get customer service"

human-to-human (Spontaneous speech): This is a bigger, more complex problem. Here we can also break it down into different applciations:

- Call Center: conversation between Agent-Customer, phone quality, compressed

- Intelligence: radio/phone/live conversations between 2 or more individuals

Now, Speech-To-Text is not what you should be saying that you care about. What you care about is solving a problem. Different technologies are used to solve different problems. See an overview here of some of them. to summarize, other approaches are Phonetic transcription, LVCSR and direct based.

Also, are you interested in being the PHd behind the technology? you would need a Masters equivalent involving *Signal processing* and probably a PHd to be cutting edge. In which case, you will work for a company that develops the actual **speech engine**. Companies like Nuance and IBM are the big ones, but also Phillips and other startups exist.

On the other hand, if you want to be the one implementing applications, you will not be working on the engine, but working on building application that USE the engine. A good analogy I think is form the gaming industry: Are you developing the graphic engine (like the Cry engine), or working on one of several hundred games, all use the same graphic engine?

Don't get me wrong, there is plenty to work on the quality of the search also outside the IBM/Nuance of the world. The engine is usually very open, and there are a lot of algorithmic tweaking to be done that can dramatically affect performance. Each business application has different constraints and cost/benefit function, so you can

make experiments for many years building better voice recognition based applications.

one more thing: in general, you would also want to have good statistics background the lower in the stack you want to be.

> At this point in time, I'm mainly interested in being able to create applications that allow automation

Good, we are converging here... Then you have no interest in "Speech-to-Text". That buzzwords takes you to the world of full transcription, a place you do not need to go to. You should be focusing on some of the more Human-to-Machine technologies like Voice XML and the ones used in IVR systems (Nuance is the biggest player there)

Share Improve this answer

Follow

edited Oct 22, 2008 at 20:56

Adam Bellaire

110k ● 19 ● 152 ● 165

answered Aug 18, 2008 at 16:56

csmba

4,083 ● 3 ● 33 ● 42

3

I would definitely recommend picking up a book or two if you are new to the field. I've got no experience in the field, so I can't make a recommendation. If you are still in

college (or still have close ties), you should find out if any of your professors can make a recommendation.

The survey you linked is probably an excellent resource, too. I'm sure there have been advancements since 1996, but the basics are unlikely to have fundamentally changed. If the survey is well-written, then it would be well worth your time to read it.

Share Improve this answer

Follow

answered Aug 18, 2008 at 16:16

Derek Park

46.8k ● 16 ● 59 ● 76

2

For OS X check out this: OS X Speech Technologies

For Windows check out this: Microsoft Speech API

Share Improve this answer

Follow

edited Aug 18, 2008 at 16:26

answered Aug 18, 2008 at 16:18

Adam Haile

31.3k ● 60 ● 195 ● 290

2

I have worked with IBMs ViaVoice product. It has a good ASR (automated speech recognition) engine, and a nice text-to-speech engine.

The websites not very good, but this is a link for the Embedded version http://www-

01.ibm.com/software/voice/support/

It is platform agnostic though, and everything works through a MVC architecture using vxml a variant of xml for voice purposes.

Share Improve this answer

Follow

answered Aug 18, 2008 at 16:55

Craig

1,396 ● 12 ● 24

0

What platform are you targeting ?. There is Microsoft Speech APIs that you can use if its for windows.

Share Improve this answer

Follow

answered Aug 18, 2008 at 16:14

rptony

1,024 ● 2 ● 12 ● 22

0

There is also the Speech Recognition Service for Android.

Share Improve this answer

Follow

answered Nov 24, 2014 at 9:47

Israel Varea

2,630 ● 2 ● 19 ● 24