Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS

Ondřej Klejch

Development of a cloud platform for automatic speech recognition

Institute of Formal and Applied Linguistics

Supervisor of the master thesis:	Mgr. Ing. Filip Jurčíček Ph.D.
Study programme:	Informatics
Specialization:	Theoretical Computer Science

Prague 2015

Dedication.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... signature of the author

Název práce: Development of a cloud platform for automatic speech recognition

Autor: Ondřej Klejch

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: Mgr. Ing. Filip Jurčíček Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt:

Klíčová slova:

Title: Development of a cloud platform for automatic speech recognition

Author: Ondřej Klejch

Department: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Ing. Filip Jurčíček Ph.D., Institute of Formal and Applied Linguistics

Abstract:

Keywords:

Contents

Introduction

The most natural form of human communication is speech. In order to be able to talk with a computer, it is crucial to have a good Automatic Speech Recognition (ASR) system. On one hand, there are several open-source ASR toolkits, however deployment of such toolkits requires substantial knowledge therefore for common software developers it is not easy to use them. On the other hand, there are a few webservices that provide ASR as a service, yet these webservices do not solve all problems - either they are paid, closed-source or they are not customizable. So **the first goal of the thesis is to develop a cloud platform for ASR** that is easy to use both from user's and maintainer's point of view.

Although accuracy of ASR systems is improving, these systems are still far from perfect. One of the reasons is that accuracy of ASR systems relies heavily on the amount of the training data and there is not enough publicly available transcribed speech data. By providing free ASR webservice it is possible to collect vast amount of recordings that can be manually transcribed and used later on for further research. Consequently, **the second goal of the thesis is to create an annotation interface** so that recordings obtained by CloudASR platform can be annotated and given back to the community.

In the following text there will be described development and deployment of CloudASR platform and of its annotation interface. Chapter 1 introduces ... In Chapter 2 architecture of CloudASR is described. Annotation interface and theory related to obtaining of human transcriptions is presented in Chapter 3. Finally, Chapter 5 concludes this thesis. User manual and programmer manual can be found in the Attachments.

1. Automatic Speech Recognition

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.1 Acoustic Models

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.2 Language Models

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.3 Decoding

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.3.1 Batch Decoding

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.3.2 Online Decoding

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.4 Transcriptions

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.5 Transcriptions via Crowdsourcing

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.5.1 Amazon Mechanical Turk

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.5.2 CrowdFlower

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.6 Open-Source ASR tools

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.6.1 HTK

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information.

Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.6.2 RWTH

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.6.3 Sphinx

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.6.4 Kaldi

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.6.5 PyKaldi

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.7 ASR cloud services

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.7.1 Google Speech API

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.7.2 Nuance

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.7.3 Tom Robinson

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.7.4 Wit.ai

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2. Used Technologies

In this chapter I will briefly present technologies that were used during development and I will explain why these technologies were used.

2.1 Platform

In the following section I will describe technologies that were used for building cloud platform. These technologies enabled us to build a scalable solution with easy deployment.

IDEA: Time consuming installation, build once use many times

IDEA: We want to make installation process as fast as possible

IDEA: Compare Docker with Virtualbox/Vagrant
Initially a virtual machine with all dependencies installed was used. Even though this satisfied all deployment requirements it does not seem useful, because this solution had much larger performance overhead and it does not allow to control the application at such fine grain level as Docker.

The most important tool used during development is **Docker** – a portable, lightweight application runtime and packaging tool[1]. It means that it allows to specify environment in which we want to run a process. The environment is described in a Dockerfile, see Figure 2.1 for an example. From a Dockerfile Docker builds an image which is later used to run that process. All dependencies are stored in the image and this image can be used on different machines. Therefore deployment time is much lower.

As a result only required dependency for running CloudASR is Docker because all other dependencies are already installed in the Docker images. Additionally, it removes bugs caused by different versions of libraries used in development and production environmnent because developers can use the same images in both environments.

IDEA: server crashes, manage all servers

TODO: cite Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center

[1] http://www.docker.com

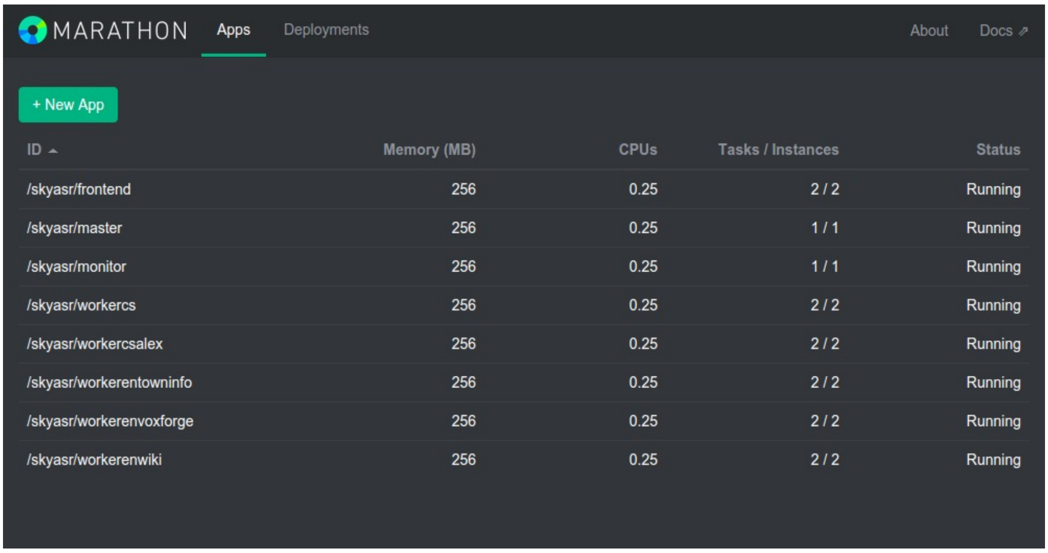
```
FROM ubuntu
MAINTAINER Ondrej Klejch

RUN sudo apt-get update && sudo apt-get install python
ADD . /opt/app
WORKDIR /opt/app

CMD python run.py
```

Figure 2.1: An example of Dockerfile.

TODO: add better caption



ID	Memory (MB)	CPUs	Tasks / Instances	Status
/skyasr/frontend	256	0.25	2 / 2	Running
/skyasr/master	256	0.25	1 / 1	Running
/skyasr/monitor	256	0.25	1 / 1	Running
/skyasr/workercs	256	0.25	2 / 2	Running
/skyasr/workercsalex	256	0.25	2 / 2	Running
/skyasr/workerentowninfo	256	0.25	2 / 2	Running
/skyasr/workerenvoxforge	256	0.25	2 / 2	Running
/skyasr/workerenwiki	256	0.25	2 / 2	Running

Figure 2.2:

TODO: An example of marathon

TODO: Describe screenshot better

The tool that allows CloudASR to run on many machines is **Mesos**. It lets users program against set of machines like it is a single machine. That means that we can create set of servers and run application on them. Also, Mesos supports Docker so images that are used in development can be also used on Mesos cluster. Furthermore, Mesos takes care about high availability of the platform. Thus, whenever some part of the CloudASR crashes Mesos will try to fix that.

Marathon[2] is a framework built on top of Mesos whose main responsibility is to launch long running applications. It also takes care about high availability (when an application crashes it tries to restart it) and easy scalability (running instances can be scaled in one click). Finally, it has a web user interface (see Figure 2.1) and REST API, through which applications can be started, scaled or stopped.

Since the traffic of CloudASR platform can be very large, it is not possible to process all HTTP requests on one machine. Therefore, there must be a load-balancer to distribute workload between running instances. CloudASR platform uses **HAProxy**[3] load-balancer, but any other load-balancers can also used with appropriate setup.

2.2 Continuos Integration & Delivery

During development of CloudASR I obeyed several practises, namely Continuous Integration and Continuous Delivery. In order to be able to do that I had to setup a platform which consisted of **Jenkins-CI**[4] and **Docker Registry**[5].

[2] https://mesosphere.github.io/marathon/

[3] http://www.haproxy.org/

[4] https://jenkins-ci.org/

[5] https://github.com/docker/docker-registry

The most important tool for Continous Integration & Delivery of CloudASR is Jenkins-CI. It watches CloudASR git repository and whenever a new code is pushed into this repository it schedules a new build of the platform. During this build the most recent code is pulled from the repository and then the docker images are built. After that tests are run to check that the new code did not break anything. Finally, successfully built images are tagged with current build number and pushed to the Docker Registry.

Docker Registry is a repository of Docker images. Even though, there are several Docker Registry providers[6], which are free for open-source projects, CloudASR uses its own free Docker Registry, in order to be also able to use proprietary software that cannot be shared with public.

2.3 Backend

The main programming language used for development is **Python**[7]. Web and REST API are built on top of **Flask**[8] microframework and they use **Gunicorn**[9] for production deployment.

IDEA: Flask - plugins, lightweight, support for SocketIO

IDEA: Gunicorn - performance, support for SocketIO...

Because CloudASR architecture consists of several nodes which need to communicate between each other. For this communication ClousASR uses **ZeroMQ**[10], because of its simple design, high performance and support for every modern language. Moreover usage of ZeroMQ makes it possible to implement each node in different programming language if needed. With ZeroMQ it is possible to create many messaging patterns, but CloudASR uses only two: request-reply and push-pull. These patterns are described in detail on Figure 2.3.

In order to be able to send complex messages via ZeroMQ sockets, messages have to be serialized. Initially, CloudASR used JSON for serialization because of its simplicity and its support in almost every language, but suddenly I found out that JSON does not support serialization of binary data, therefore, I had to choose another serialization format. As a result, CloudASR uses **Google Protocol Buffers**[11], which has support in many languages, allows specification of various message types (See Figure 2.3 for example) and serializes messages in very compact way (See Table 2.3 for a comparison of different serializations).

[6] https://hub.docker.com/, https://quay.io/

[7] https://www.python.org/

[8] http://flask.pocoo.org/

[9] http://gunicorn.org/

[10] http://zeromq.org/

[11] https://developers.google.com/protocol-buffers/

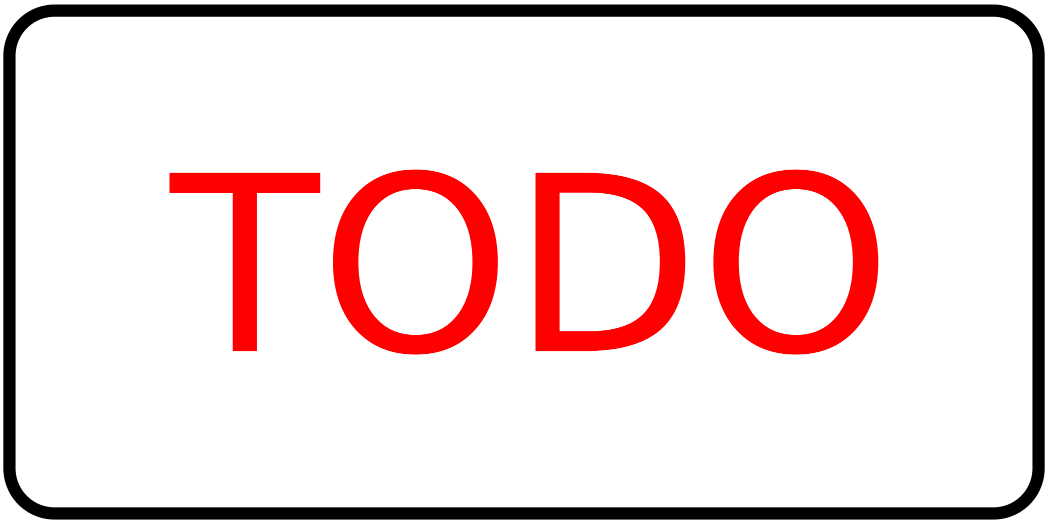


Figure 2.3: Description of used ZeroMQ patterns.

raw file size	56146	
bytes_protobuf	56118	0.999x
base64	74872	1.333x
json_array	158590	2.824x

Table 2.1:

TODO: add better description

2.4 Frontend

Frontend uses several well-known open-source libraries, namely, **Twitter Bootstrap**[12] for CSS styling of the web, **jQuery**[13] and **Angular.js**[14] for interactive elements on the web.

Modern web browsers supports **WebAudio API**[15], which is a high-level JavaScript API for processing and synthesizing audio in web applications. One of the things that can be done with this API is recording of an audio. Thus, it is possible to create a web demo for CloudASR online speech recogniser. The demo is based on **Recorder.js**[16] library, which can record output of WebAudio API and return it as a PCM chunks.

Next step is to send these chunks to the API. Because the demo demonstrates the online speech recognition mode, it is not possible to wait for whole recording to be recorded and then send it to the API via HTTP POST request, thus, CloudASR uses **Socket.IO**[17] to send stream of chunks to the API and to receieve stream of results from the API.

[12] `http://getbootstrap.com/2.3.2/`

[13] `https://jquery.com/`

[14] `https://angularjs.org/`

[15] `http://webaudio.github.io/web-audio-api/`

[16] `https://github.com/mattdiamond/Recorderjs`

[17] `http://socket.io/`

```
message HeartbeatMessage {
  required string address = 1;
  required string model = 2;
  required Status status = 3;

  enum Status {
    STARTED = 0;
    WAITING = 1;
    WORKING = 2;
    FINISHED = 3;
  };
}
```

Figure 2.4: An example of Google Protocol Buffer message specification.

TODO: Add beter caption

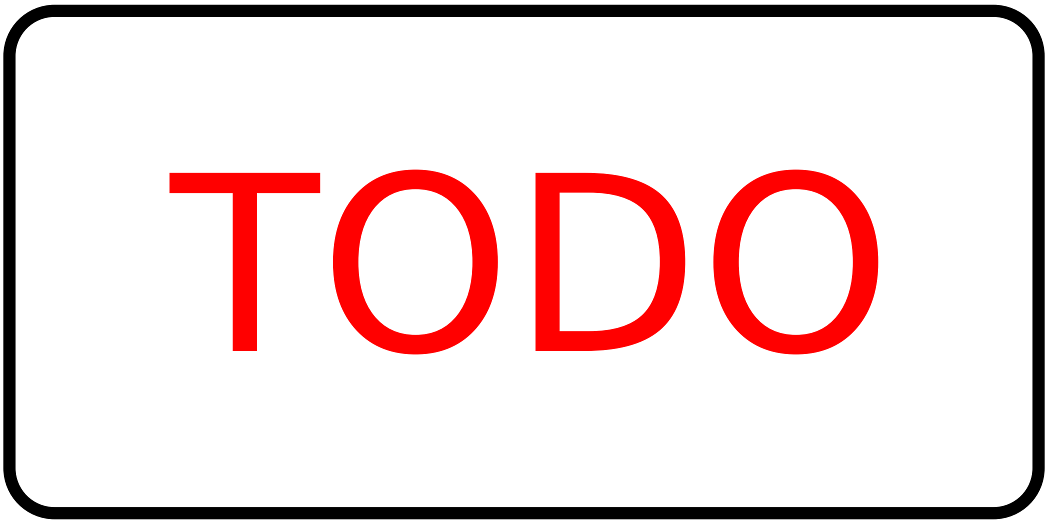


Figure 2.5: CloudASR Web Demo

3. Solution

In this chapter I will describe the implementation of CloudASR, I will stress key design choices...

3.1 Architecture

The key requirement for the architecture of CloudASR is scalability, because speech recognition is a demanding task and it is not possible to handle many parallel requests on a single machine. Thus, the platform was designed from the very beginning to be able to run on many machines.

IDEA: Distributed architecture

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

IDEA: Message/Actor

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

IDEA: Queue vs. Sockets.

IDEA: Dependency Injection—Factory Methods—Testability

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1.1 API

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and

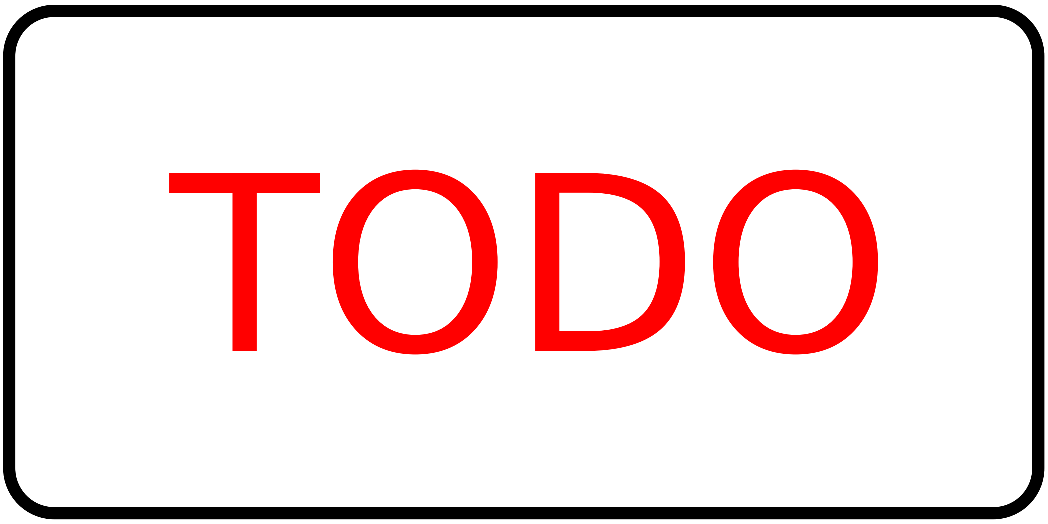


Figure 3.1: Architecture

some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1.2 Web

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1.3 Master

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1.4 Worker

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information.

Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1.5 Recordings saver

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.2 Request Workflow

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.2.1 Batch Recognition

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.2.2 Online Recognition

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet

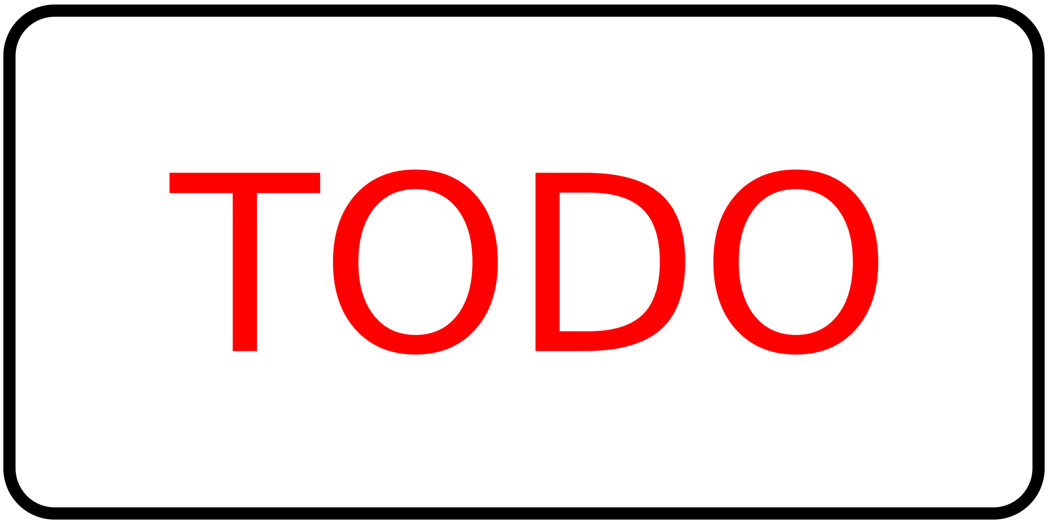


Figure 3.2: Batch Workflow

and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

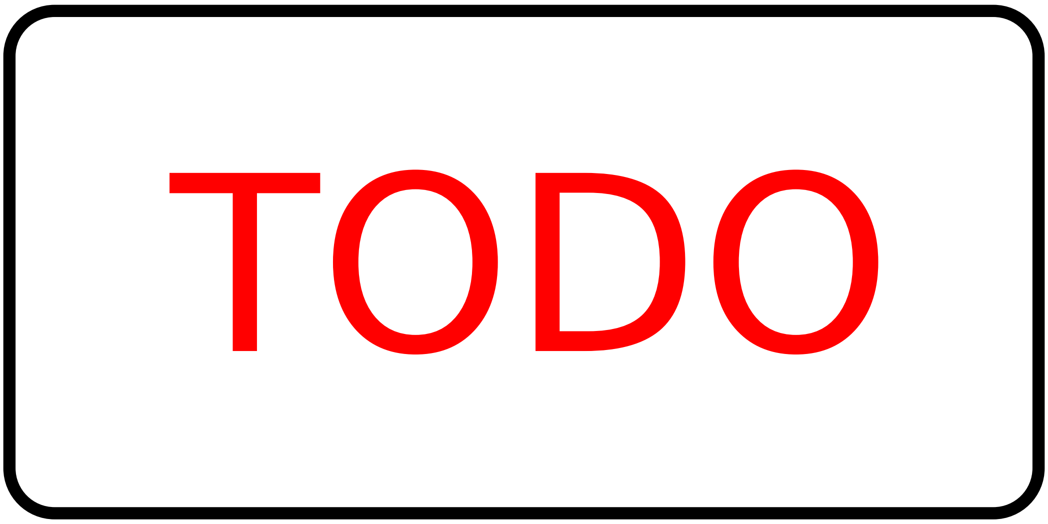


Figure 3.3: Online Workflow

3.3 Deployment

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.3.1 Single-Host Deployment

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.3.2 Multi-Host Deployment

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.3.3 Scalability

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.3.4 Contiunous Integration & Countinuous Delivery

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.4 Hosting Various Models

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information.

Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.4.1 Worker Deployment

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.4.2 Deployment of New Kaldi Worker

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.4.3 Deployment of Arbitrary Worker

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4. Evaluation

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.1 CloudASR Platform Benchmarks

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.2 Batch Recognition Benchmark

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in

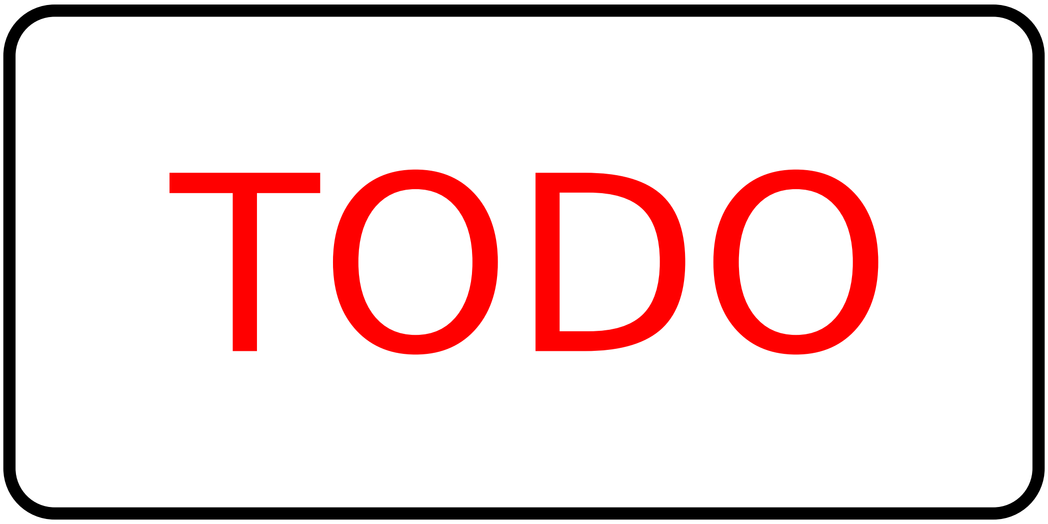


Figure 4.1: fig:batch-benchmark

of the original language. There is no need for special content, but the length of words should match the language.

4.3 Online Recognition Benchmark

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like

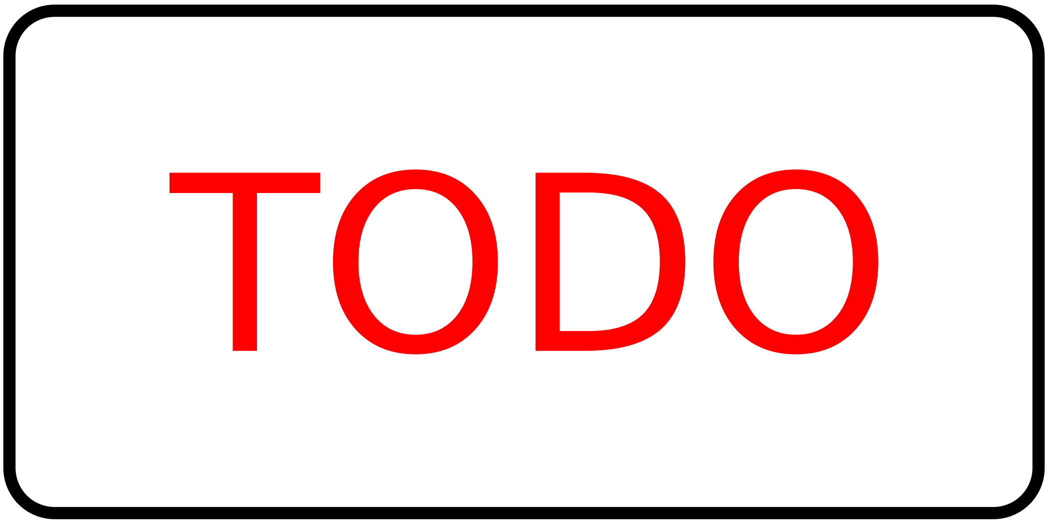


Figure 4.2: fig:online-benchmark

at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Conclusion

Goals of this thesis were to develop a cloud platform for ASR, CloudASR, and an annotation interface for annotating speech data. These goals were successfully accomplished and in several aspects even surpassed - in addition to original requirement to create batch recognition mode, we also implemented online recognition mode. In the following sections we summarize our achievements in detail and at the end we propose ideas for future work.

Cloud platform for ASR

The first goal of this thesis was to develop a cloud platform for ASR, CloudASR, that would provide batch API for speech recognition of wave files. The platform uses Master/Worker architecture. Consequently, it is able to run both on single-machine and multi-machine setup. The platform allows us to run workers for various language models and to scale workers according to our needs. To be able to run CloudASR on several machines we chose Mesos/Marathon as an underlying technology The current implementation of the API supports two modes of speech recognition: batch and online.

Firstly, batch mode allows users to send a file with a recording to the server and then it sends transcribed text back as a json. API of this mode is similiar to Google Speech API which allows users to switch from Google Speech API to CloudASR easily.

Secondly, users can transcribe speech recordings in real-time via online mode. We have also created Python and JavaScript libraries for using our API. JavaScript library achieves similiar latency as WebkitSpeechRecognition in Google Chrome

TODO: add benchmark

.

Finally, we wanted CloudASR to be easily deployable. Because of that, we used Docker for creating and running application containers. As a result only dependency that users have to install is Docker for single-node setup and Mesos Cluster for multi-node setup. Moreover, installation scripts for these dependencies are included within the distribution together with deployment scripts, that can be used for CloudASR instances management.

Annotation interface

The second goal of this thesis was to create an annotation interface for annotating speech data. First responsibility of the annotation interface is to collect and store obtained recordings.

The second responsibility is to allow users to rate transcriptions of the recordings (Is the transcription correct? yes/no) or to subsequently add their own transcriptions. The annotation interface implements algorithm to choose golden transcription from several manual transcriptions that were obtained for the recording. Additionally it is also possible to add manual transcriptions via external job at CrowdFlower.

The third responsibility is to provide export of transcribed recordings. This can be done either by downloading archive from the web or by using Torrent.

Future work

- Since manual transcription of recordings is expensive it would be good to make users transcribe only parts of the recordings in which ASR system wasn't confident enough

  TODO: cite (http://www.phontron.com/paper/sperber14slt.pdf)

  . This idea could be used for both user transcription and CrowdFlower transcription.

- With manually transcribed recordings from CloudASR platform it is possible to continuously improve accuracy of the underlying ASR system by adapting the language model to the type of language that the users of the CloudASR really use. Thus CloudASR could provide an option to automatically update language model when a certain amount of new transcribed recordings was collected.

- Because running CloudASR platform is expensive in terms of costs for a server hosting, it would be good to optimize usage of individual workers so that spare workers are shut down when there is no need for them and new workers are started when the traffic arise. This can be achieved either by providing feedback control based systems

  TODO: cite (http://shop.oreilly.com/product/0636920028970.do)

  or by using machine learning techniques.

  TODO: cite

- As CloudASR platform provides API for speech recognition, it could also be used for another speech related tasks like Language Identification, Speaker Identification, Voice Activity Detection, etc.

Bibliography

List of Tables

List of Abbreviations

Attachments