Speech recognizer for EllaVator

Arif Khan

Saarland University

arifkhan@coli.uni-saarland.de

Wednesday 3rd June, 2015

Overview

1. Speech recognizer - background
   - Components of speech recognizer
   - Acoustic model (AM)
   - Language model (LM)
     - Grammar based LM
     - Statistical LM
2. Using LM with opendial
3. Training Models
   - Installing Sphinx for training
   - Training acoustic model

Components of speech recognizer

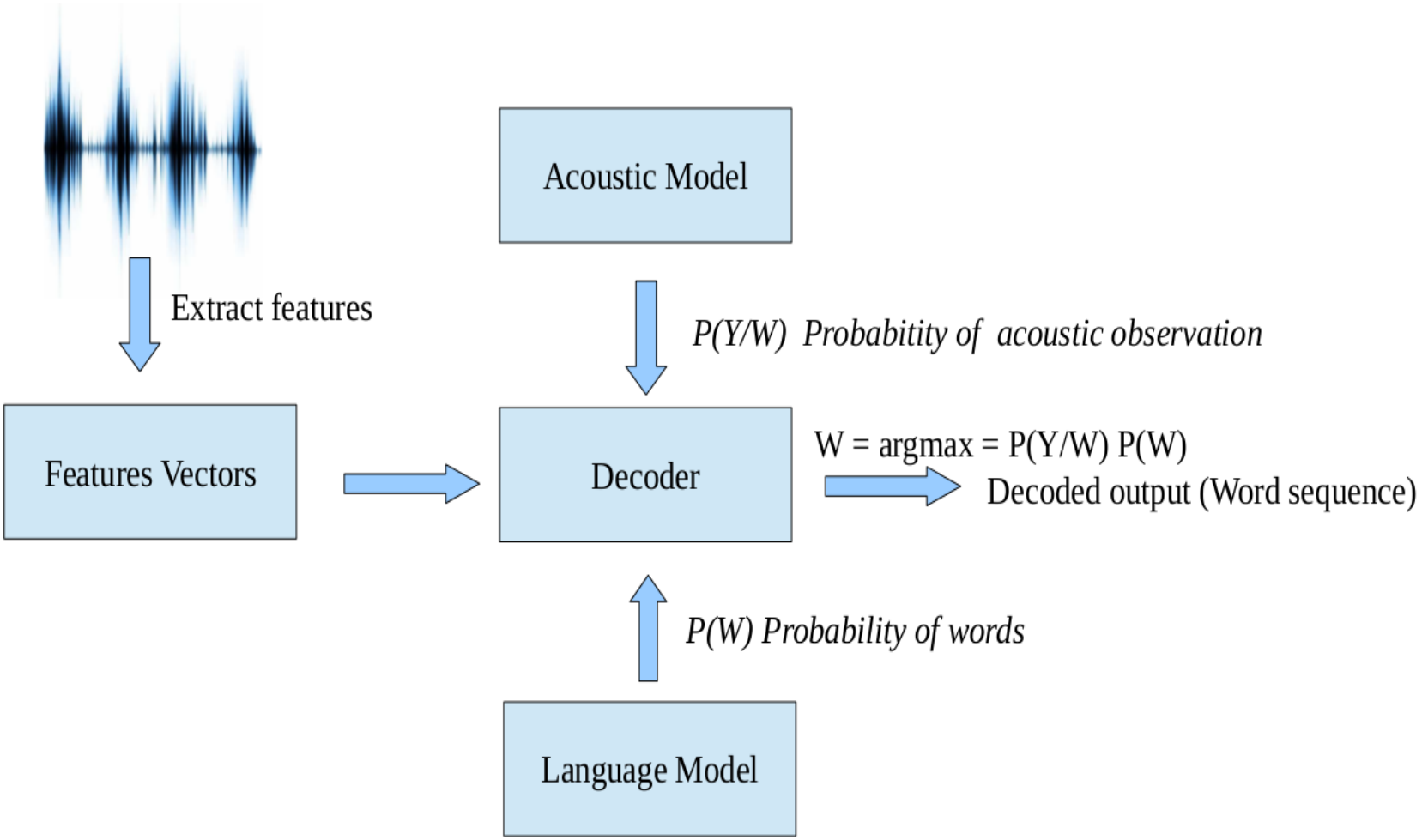
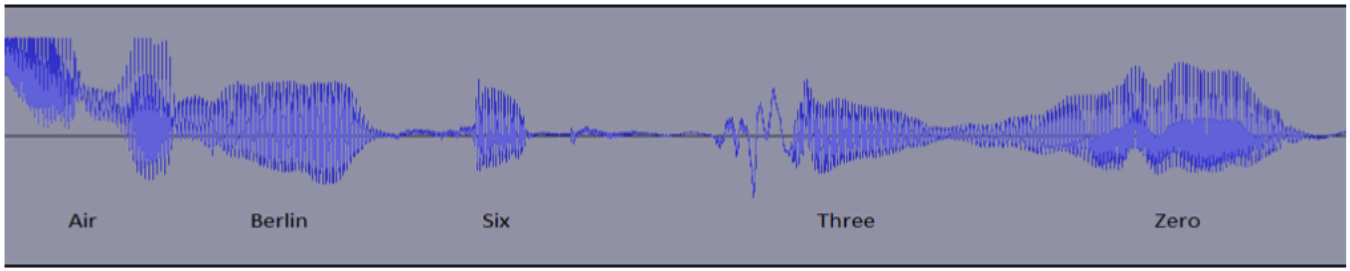


Figure: *Main components of speech recognizer*

Acoustic model - AM



- Pronouncing dictionary:

AIR_BERLIN	EH1 R B ER0 L IH1 N
SIX	S IH1 K S
THREE	T R IY1
ZERO	Z IH1 R OW0

- ~39-45 phonemes are used for English Language

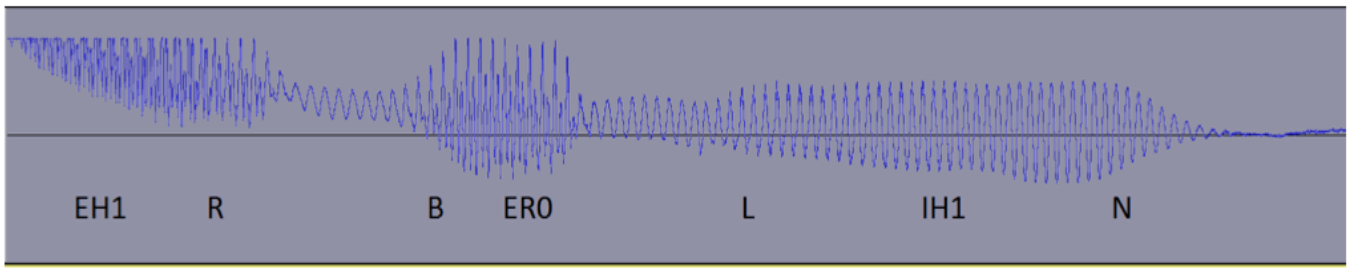


Figure: *Acoustic model*

Language model

Grammar based LM

- Write grammars to specify the possible sentence structures
- Good for small set of sentences (small domains)
- In some specification, weights can be assigned to sentence structure

- With Sphinx plugin of opendial, JSpeech Grammar Format (JSGF) is used for specifying grammar
- Provides a lot of flexibility for writing grammar (enough for EllaVator)
- see documentation for complete specification and examples.
- JSGF also inherits the drawback of grammar based LM

Language Model

Example

grammar ellavator;
public <ellavator> = <begin> | <command>;
<begin> = (Hello | “Good morning” | “Good evening”) Ella;
<command> = (<start> <floor_no>);
<start> = [Take me to] | [please];
<floor_no> = (first | second | third | fourth) floor;

Language Model

Statistical LM

- Statistical LM gives probabilistic estimates of word strings from large text corpora of transcribed speech.
- The probabilities for a word are approximated from the preceding sequence.
- The preceding sequence could be one (bigram), two (trigram) words
- For trigrams we have:

$$P(w_k|w_{k-1}, w_{k-2}) = \frac{count(w_{k-2}, w_{k-1}, w_k)}{total(w_{k-2}, w_{k-1})} \quad (1)$$

Statistical LM - Example

<001> Hello Ella fourth floor
<002> Good morning Ella fourth floor
<003> Good evening Ella four floor

For probability of "Ella" if "Hello" is already spoken:

$$P(Ella|Hello) = \frac{count(Ella, Hello)}{total(Hello)} \tag{2}$$

Language Model

Statistical LM - Example

- With Sphinx we can also use statistical LM trained by various tools.
- Some tools that we can use for training are: cmulmtk, IRSLM, MITLM, SRILM, `http://cmusphinx.sourceforge.net/wiki/tutoriallm`
- we can also use the online interface of cmulmtk for small set of sentences. `http://www.speech.cs.cmu.edu/tools/lmtool-new.html`
- Dont forget to pre-process the data before using the online tool (web interface)

Language Model

Statistical LM - Example

```
\3-grams:
-0.6021 <s> GOOD EVENING
-0.6021 <s> GOOD MORNING
-0.3010 <s> HELLO ELLA
-0.3010 ELLA FOUR FLOOR
-0.3010 ELLA FOURTH FLOOR
-0.3010 EVENING ELLA FOUR
-0.3010 FOUR FLOOR </s>
-0.3010 FOURTH FLOOR </s>
-0.3010 GOOD EVENING ELLA
-0.3010 GOOD MORNING ELLA
-0.3010 HELLO ELLA FOURTH
-0.3010 MORNING ELLA FOURTH

\end\
```

Figure: *Bigram language model for Ella*

Language Model

Which one is good

- Statistical LM captures the actual probabilities from corpus.
- Weights can be assigned to grammar based LM, but weights are static.
- Empty loops are valid sentences in grammar based LM, if the grammar is poorly written.

Using LM with opendial

Using LM with opendial

- Write a grammar for EllaVator in SJGF that covers the user utterances.
- Come up with example utterances using the grammar you wrote

Installing Sphinx for training acoustic model

Downloading the following

- sphinxbase, sphinxtrain, pocketsphinx from `https://github.com/cmusphinx`
- Tutorial for training acoustic model `http://cmusphinx.sourceforge.net/wiki/tutorialam`