Evaluation of speech technologies

CLARA Training course on evaluation of Human Language Technologies

Djamel Mostefa

Evaluations and Language resources Distribution Agency

November 27, 2012

Evaluation of speech technologies

Evaluation of speaker recognition

Evaluation of speech recognition

Evaluation of spoken language understanding

Evaluation of speech synthesis technologies

Concrete examples

Evaluation of speaker identification

Evaluation of speech recognition

Speech technologies

- Speaker recognition

Speech technologies

- Speaker recognition
- Automatic speech recognition

Speech technologies

- Speaker recognition
- Automatic speech recognition
- Spoken language understanding

Speech technologies

- Speaker recognition
- Automatic speech recognition
- Spoken language understanding
- Speech synthesis

Evaluation of speaker recognition

Automatic speaker recognition technologies try to recognize a person based on a speech sample. Speaker recognition covers two sub areas: speaker identification and speaker verification.

- Speaker identification: the goal is to classify an unlabeled voice token as belonging to a set of known speakers.
    - if all trials are coming from the set of know speakers, the task is a closed set speaker identification.
    - if trials can come from speakers out of the list of known speakers, the task is an open set speaker identification
- Speaker verification : the task is to decide whether or not the unlabeled voice belongs to a claimed identity.

We can distinguish text *dependent* speaker recognition to text *independent* speaker recognition.

How do we evaluate speaker recognition ?

How do we evaluate speaker recognition ?

- Define the task (speaker identification vs verification, matched vs unmatched conditions, text (in)dependent, . . .)

How do we evaluate speaker recognition ?

- Define the task (speaker identification vs verification, matched vs unmatched conditions, text (in)dependent, . . .)
- Produce corpora for evaluation (train, development, <u>test</u>)

How do we evaluate speaker recognition ?

- Define the task (speaker identification vs verification, matched vs unmatched conditions, text (in)dependent, . . .)
- Produce corpora for evaluation (train, development, test)
- Evaluate the performance by using different metrics

Metrics for the evaluation of speaker identification

- For the closed set speaker identification, the metric is called the Misclassification Rate (M) as:

$$M = \frac{N_{err}}{N_{tot}}$$

where N_{err} is the number of trials in error and N_{tot}, the total number of trials.

Metric for the evaluation of speaker identification

- For the open set identification, we have two metrics, the Misclassification Rate as defined previously and the False Rejection Rate or Miss for trials of speakers of the closed set rejected wrongly.

$$E_{miss} = \frac{N_{miss}}{N_{target}}$$

where N_{miss} is the number of trials of speakers of the closed set not detected and N_{target}, the total number of target trials.

Metrics for the evaluation of speaker verification

Speaker verification is a detection task for which a claimed identity has to be confirmed or declined based on a voice sample. Two kinds of error can be observed, false acceptation when an impostor is accepted as the speaker he claimed to be and false rejection when a correct speaker is rejected.

- False Rejection Rate or Miss as defined previously:.

$$E_{miss} = \frac{N_{miss}}{N_{target}}$$

- False Acceptation Rate :

$$E_{far} = \frac{N_{fa}}{N_{impostors}}$$

where N_{fa} is the number of trials falsely detected as the target speaker and $N_{impostors}$ the total number of impostor trials.

Metrics for the evaluation of speaker verification

Evaluating speaker verification is done with to metrics but usually we try to have only one number. So there are different solutions to combine False rejection and False acceptation:

- Equal Error Rate: using the decision threshold for which $E_{miss} = E_{far}$
- Geometric Mean Error Rate:

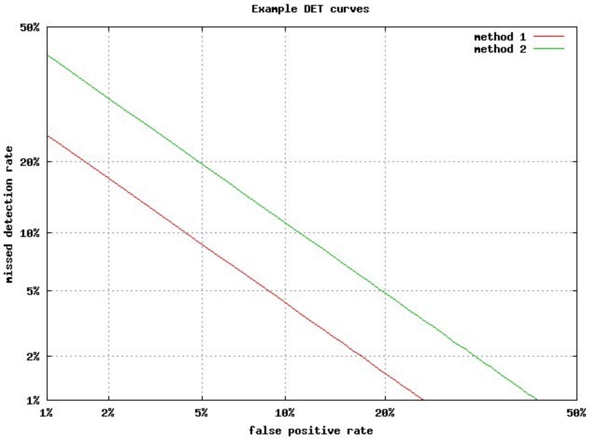
$$GME = \sqrt{E_{miss} * E_{far}}$$

- Cost function :

$$cost = c_{miss} * E_{miss} * P_{target} + c_{fa} * E_{far} * (1 - P_{target})$$

- Detection Error Trade-off plot (DET curve): it's a plot of miss probability as a function of false acceptation probability.

Example of a DET curve



Evaluation of speech recognition

Automatic speech recognition (ASR) or speech-to-text is the process of transcribing automatically audio data. It includes various technologies like:

- Voice commands (navigation systems, mobile devices, . . .)
- Telephony speech recognition (Interactive Voice Response server)
- Voicemail transcription
- Transcription of broadcast news
- Conversational speech recognition
- Meeting transcription
- Dictation
- . . .

Metric for the evaluation of ASR

Evaluation of ASR systems is performed by comparing or aligning a reference transcription to an hypothesis transcription (system output). The most commonly used metric for assessing the performance of ASR systems is the Word Error Rate (WER):

$$WER = \frac{N_{ins} + N_{sub} + N_{del}}{N_{tot}}$$

where

- N_{ins} is the number of inserted words in the hypothesis
- N_{sub} is the number of substituted words in the hypothesis
- N_{del} is the number of deleted words int the hypothesis
- N_{tot} is the total number of words int the reference

Concretely how do we evaluate the speech recognition systems?

- Define the task: language, domain, processing and time constraints, . . .
- Produce corpora for evaluation
  - training data of 100h of transcribed speech or more
  - development of 3 to 6 hours, (30k to 60k words)
  - test of 3 to 6 hours
- Evaluate the performance with WER

A concrete example: evaluation of broadcast news transcription systems

- Record or acquire audio data (for example broadcast news shows)
- Manually transcribe the data orthographically with specific tools like Xtrans or Transcriber
- Convert and normalize the transcription to an evaluation format called STM file
- Evaluate the performance by aligning an hypothesis CTM file with a reference STM file

reference STM file format

It is a plain text file format with 7 columns

- audiofilename
- channel
- speaker_name
- starttime
- endtime
- speakergenre
- transcription

Example of STM file

```
20041006_0700_0800_CLASSIQUE 1 Emmanuel_Cugny 8.552 11.702 (o,f0,male) mais tout de suite les grands titres
de l' actualité maude bayeu bonjour
20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 11.702 18.116 (o,fx,female) bonjour le gouvernement et l'
opposition enterrent la polémique sur la mission de didier julia pour libérer les otages en irak
20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 18.116 23.329 (o,f3,female) aprés une réunion  matignon hier
ils ont décidé de rejouer la carte de la cohésion sociale
20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 23.329 27.144 (o,f3,female) du moins dans l' attente de la
libération de christian chesnot et georges malbrunot
20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 27.144 31.080 (o,f3,female) le premier ministre n' en a pas
moins condamn? l' initiative paralléle du déput?
20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 31.080 37.554 (o,f3,female) ump celui ci s' est expliqué hier
huis clos devant la commission des affaires étrangéres de l' assemblée nationale
20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 37.554 38.709 (o,f3,female )le groupe ump

20041006_0700_0800_CLASSIQUE 1 Maude_Bayeu 77.795 80.698 (o,f3,female) (%HESITATION) dick cheney a

justifié l' intervention dans le pays
```

hypothesis CTM file format

It is a plain text file format with 5 columns

- audiofilename
- channel
- starttime
- duration
- recognized word

Example of CTM file

```
20041006_0700_0800_CLASSIQUE 1 11.12 0.55 bonjour
20041006_0700_0800_CLASSIQUE 1 11.88 0.52 bonjour
20041006_0700_0800_CLASSIQUE 1 12.44 0.21 le
20041006_0700_0800_CLASSIQUE 1 12.65 0.62 gouvernement
20041006_0700_0800_CLASSIQUE 1 13.27 0.04 et
20041006_0700_0800_CLASSIQUE 1 13.31 0.05 l'
20041006_0700_0800_CLASSIQUE 1 13.36 0.63 opposition
20041006_0700_0800_CLASSIQUE 1 13.99 0.38 enterre
20041006_0700_0800_CLASSIQUE 1 14.37 0.11 la
20041006_0700_0800_CLASSIQUE 1 14.48 0.51 polémique
20041006_0700_0800_CLASSIQUE 1 14.99 0.18 sur
20041006_0700_0800_CLASSIQUE 1 15.17 0.10 la
20041006_0700_0800_CLASSIQUE 1 15.27 0.37 mission
20041006_0700_0800_CLASSIQUE 1 15.64 0.17 de
20041006_0700_0800_CLASSIQUE 1 15.81 0.36 Didier
20041006_0700_0800_CLASSIQUE 1 16.17 0.33 Julia

20041006_0700_0800_CLASSIQUE 1 16.50 0.14 pour
```

How do we score the system hypothesis practically ?

- We are using a scoring software developed by NIST (National Institute of Standards and Technology) called `sclite`
- We simply run the following command:
- sclite -D -F -r reference.stm stm -h fhyphothesis.ctm ctm
- sclite first aligns the sequence of words of the hypothesis sentence with the sequence of words of the reference by using the Levenshtein distance also called the "edit distance"

Some alignment examples

Some alignment examples

- REF sept heures et quart maude bayeu nous rappelle les grands titres de l' actualité

Some alignment examples

- REF sept heures et quart maude bayeu nous rappelle les grands titres de l' actualité
- HYP ou sept heures et quart monde bayeux nous rappelle les grands titres de l' actu lit et

Some alignment examples

- REF sept heures et quart maude bayeu nous rappelle les grands titres de l' actualité
- HYP ou sept heures et quart monde bayeux nous rappelle les grands titres de l' actu lit et
- REF: ** sept heures et quart MAUDE BAYEU nous rappelle les grands titres de l' **** *** ACTUALIT

  HYP: OU sept heures et quart MONDE BAYEUX nous rappelle les grands titres de l' ACTU LIT ET

Some alignment examples

- REF sept heures et quart maude bayeu nous rappelle les grands titres de l' actualité
- HYP ou sept heures et quart monde bayeux nous rappelle les grands titres de l' actu lit et
- REF: ** sept heures et quart MAUDE BAYEU nous rappelle les grands titres de l' **** *** ACTUALIT

  HYP: OU sept heures et quart MONDE BAYEUX nous rappelle les grands titres de l' ACTU LIT ET
- Eval: I S S I I S

Evaluation of spoken language understanding

- Spoken language understanding systems goal is to interpret what has been said
- It relies on a semantic representation of a speech file
- The semantic knowledge is represented by the definition of an ontology and semantic concepts

Evaluation of spoken language understanding

In a project called MEDIA of evaluation of spoken language understanding systems, we defined semantic segments as a 3-tuple which contains

- the mode, affirmative or negative
- the name of the concept representing the meaning of the sequence of words
- the value of the concept

The evaluation can be done by computing the concept error rate CER (similar to the WER defined previously).

Evaluation of speech synthesis technologies

The evaluation of speech synthesis systems is mainly subjective by asking listeners to evaluate the quality of the audio output based on Mean Opinion Score (MOS) MOS uses a 5 scale:

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

In addition to the subjective evaluation, there are also automatic evaluation to measure the performance of grapheme-to-phoneme converter. We compute the phone error rate (similar to word error rate)

Login information

- Connect to the server using the following command:
- ssh -l LOGIN 192.168.1.220
- with the following login/passwords:
- 10 different accounts LOGIN: clara1 / clara2 / .../ clara10
- password: clara for the 10 accounts

Concrete case of evaluation of speaker identification

Concrete case of evaluation of speaker identification

- This was an evaluation campaign of speaker identification organized with the EU CHIL project.

Concrete case of evaluation of speaker identification

- This was an evaluation campaign of speaker identification organized with the EU CHIL project.
- Log in to the server and go the ACOUSTIC_PERSON_IDENTIFICATION directory

Concrete case of evaluation of speaker identification

- This was an evaluation campaign of speaker identification organized with the EU CHIL project.
- Log in to the server and go the ACOUSTIC_PERSON_IDENTIFICATION directory
- The exercise is to evaluate the Misclassification Rate of the different submissions from AIT, CMU, LIMSI, MIT, UIUC, UPC,

Concrete case of evaluation of speaker identification

- This was an evaluation campaign of speaker identification organized with the EU CHIL project.
- Log in to the server and go the ACOUSTIC_PERSON_IDENTIFICATION directory
- The exercise is to evaluate the Misclassification Rate of the different submissions from AIT, CMU, LIMSI, MIT, UIUC, UPC,
- What are the performances ?

Concrete case of evaluation of speech recognition

Concrete case of evaluation of speech recognition

- This was a French evaluation campaign of broadcast news called ESTER

Concrete case of evaluation of speech recognition

- This was a French evaluation campaign of broadcast news called ESTER
- The exercise is to evaluate the Word Error Rate of the different submissions from LIA, LIMSI and LIUM

Concrete case of evaluation of speech recognition

- This was a French evaluation campaign of broadcast news called ESTER
- The exercise is to evaluate the Word Error Rate of the different submissions from LIA, LIMSI and LIUM
- What is the WER for each system ?