COMBINING CRITERIA FOR THE DETECTION OF INCORRECT ENTRIES OF NON-NATIVE SPEECH IN THE CONTEXT OF FOREIGN LANGUAGE LEARNING

Luiza Orosanu, Denis Jouvet, Dominique Fohr, Irina Illina , Anne Bonneau
Speech Group, LORIA (Inria, CNRS, Universite de Lorraine)
Nancy, France

Introduction

- How does an automatic system for foreign language learning work?
  - the system displays a word or a sentence on the screen
  - the learner must pronounce and record the expected sentence
  - the system analyzes the acoustic signal that has just been recorded
  - the learner receives the feed-back on the quality of his pronunciation
- What could go wrong?
  - the learner could be distracted by the environment
  - the learner might pronounce a different sentence, or skip a few words
  - a technical problem might appear during the recording
- **What is our objective?**
  - introduce a detector of incorrect entries before starting the analysis
  - make sure that the received data can be considered as being "correct"

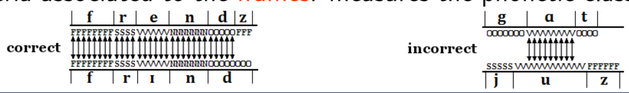
Decode the audio signals in three different ways

- Constrained decoding: the system is forced to follow the sequence of words within the expected text
- Phonetic decoding based on phoneme loop: the system is free to choose any phoneme in any position in the sentence
- Phonetic decoding based on word loop: the system is free to choose any word in any position in the sentence

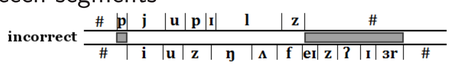
Compare a constrained decoding with an unconstrained one

- Comparison criteria associated to the phonemes: measures the phonemes adequacy

- Comparison criteria associated to the frames: measures the phonetic class adequacy

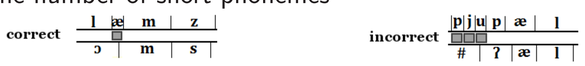


- Comparison criteria associated to the non-speech segments: measures the duration difference of non-speech segments



- Comparison criteria associated to the log likelihood ratio: measures the difference between the logarithmic likelihoods

- Comparison criteria associated to the phonemes of minimal duration: measures the difference between the number of short phonemes



Entry classifier

- Define the training data set $D = \{\bar{X}_i, y_i\}, i = 1, ..., N$ where:
  - $\bar{X}_i = \{x_1, x_2, ..., x_k\}$ is the vector containing **k** comparison criteria
  - $y_i = 1$ (correct entry) or 0 (incorrect entry)
  - $N =$ the number of entries within the training data set

- Compute an entry's probability of being correct (logistic regression)

$$f(\bar{X}) = \frac{1}{1+\exp(-(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+...+\alpha_k x_k))}$$

- Estimate the α parameters by minimizing the error function

$$E = -\sum_{i=1}^{N} \left(y_i \cdot \ln(f(\bar{X}_i)) + (1 - y_i) \cdot \ln(1 - f(\bar{X}_i)\right)$$

- Evaluate the classifier's performance
  - compute error rates for various values of a $0 \leq \sigma \leq 1$ threshold
  - if $f(\bar{X}) > \sigma$ then the entry represented by the $\bar{X}$ criteria is accepted
  - False Acceptance $FA = \frac{\text{incorrect entries wrongly rejected}}{\text{incorrect entries}}$
  - False Rejection $FR = \frac{\text{correct entries wrongly rejected}}{\text{correct entries}}$
  - F-measure $\frac{1}{F} = \frac{1}{2}\left(\frac{1}{1-FA} + \frac{1}{1-FR}\right)$

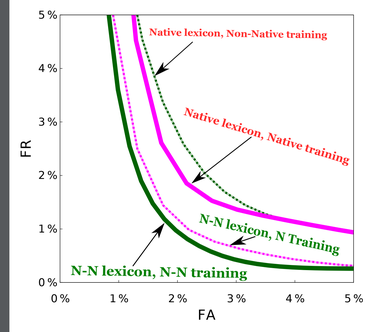
Experimental setup

- Non-native corpora
  - INTONALE Project
  - ∼ 800 English sentences
  - 34 French speakers (29 women, 5 men)
  - 50% for training, 50% for testing (results displayed on poster)

- Native corpora
  - INTONALE Project
  - ∼ 1500 English sentences
  - 22 English speakers (15 women, 7 men)
  - 50% for training, 50% for testing (results presented in the paper)

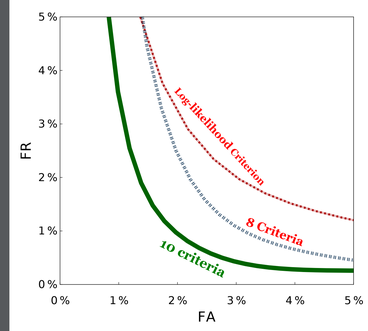
- HMM toolkit: HTK
- Acoustic features: MFCC (12 MFCC coefficients + temporal derivates + the logarithm of the energy per frame)
- Acoustic models: HMM (16 gaussian mixtures)
- Two lexicons:
  - native (CMU)
  - non-native (includes non-native variants)

Impact of the lexicon and the training data set



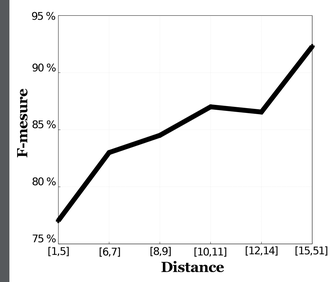
- non-native lexicon
- non-native training data set

Impact of the comparison criteria



- comparison of the forced alignment with both phoneme loop and word loop alignments
- all 10 comparison criteria (5 criteria per comparison)

Overall performance



- distance: measures the difference between the original correct transcription (which should be accepted) and the modified transcription (which should be rejected)
- the F-measure gets greater than 80% when difference over 6 phonemes

Conclusions

- Our experiments have shown that it is important to:
  - train the decision function on non-native data
  - use non-native pronunciations in the lexicon
  - combine all 10 comparison criteria
- The optimal setting leads to a classifier able to detect incorrect entries when more than 6 phonemes are wrong

ALLEGRO Project, funded by the European program INTERREG IV.

{firstName.lastName}@loria.fr