Hybrid language models for speech transcription

Luiza Orosanu
Denis Jouvet

Speech Group, Loria
Nancy, France

Introduction

- **Main objective of the RAPSODIE project**
  - automatic speech transcription
    - adapted to the needs of deaf or hard of hearing people
      - improve communication between deaf people and their entourage
      - tool of socialization and/or integration in the workplace
    - under real-time operating constraints
      - limited memory & computing power for possible embedded solution
- **Approach**
  - target only people with a good knowledge of written French
  - optimization of recognition models (and display format) for this task

Extracting relevant linguistic information

- previous work has compared different linguistic units for phonetic decoding: words, phonemes, syllables → syllables offer a good performance
- interviews with deaf people has emphasized the importance of words for understanding the message
- whatever the vocabulary size is, out-of-vocabulary words occur
- compromise: combine words and syllables into a single language model
  - ensure proper recognition of the most frequent words
  - provide sequences of syllables for the speech segments out-of-vocabulary

Settings

- Configuration
  - MFCC acoustic analysis : 32 ms window, 10 ms shift → 12 MFCC parameters and the logarithm of the energy per frame ($+\Delta, \Delta\Delta$)
  - SRILM for training the language models
  - Sphinx3 for training the gender dependent HMM acoustic models (with 64 Gaussian component mixtures)
  - PocketSphinx for speech decoding and confidence measure computation (posterior probability)
- Data
  - For training the **phonetic acoustic models**
    - training sets of ESTER2 and ETAPE & transcribed data of EPAC
    - about 300 hours of speech and 4 million words
  - For training the **hybrid language models**
    - training sets of ESTER2, ETAPE et EPAC **after a forced alignment and transformation into hybrid unit sequences** (words+syllables)
  - For performance evaluation: development sets of ESTER2 and ETAPE

Creating a hybrid language model

- establish a training corpus based on hybrid lexical features
- define the lexicon vocabulary by choosing
  - the most frequent words
  - the syllables corresponding to out-of-vocabulary words
- **Method to define the syllables**
  - training corpus fully phonetized (by forced alignment)
    - to take into account the **'liaison' & reduction** events
  - sequence of phonemes treated by a syllabification tool
  - syllabification rules **[Bigi et al, 2010]**
    - a syllable contains a single vowel
    - a pause designates a syllable's boundary
    - rules specify the syllable boundary for sequences of phonemes, as for example:

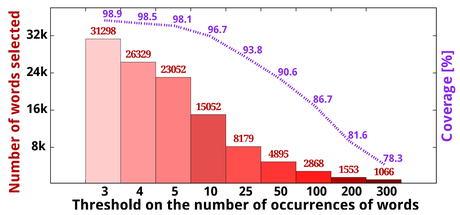
Sequence of phonemes	Split position	Resulting syllables
VV	0	V V
VxV	0	V xV
VxxV	1	Vx xV

- **Example of a "words & syllables" transcription**

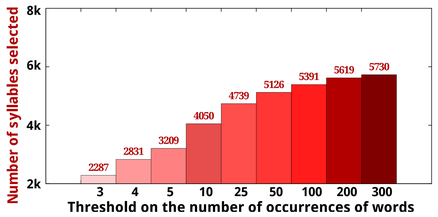
quel est le prix du **tournevis**
quel est le prix du **t u r n swa v i s** ← forced alignment
quel est le prix du **t_u_r n_swa v_i_s** ← words & syllables

- **according to different minimum thresholds** on the frequency of occurrence of words: $\theta \in \{3, 4, 5, 10, 25, ...\}$
  → different transcriptions of the training corpus
  → different lexicons and language models

- How many words are modeled inside the hybrid LM?

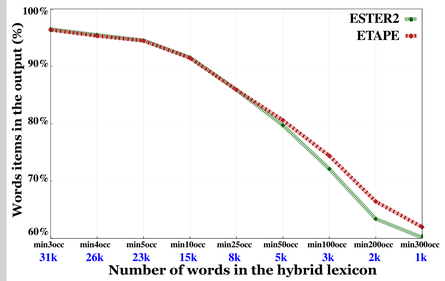


- How many syllables are modeled inside the hybrid LM?

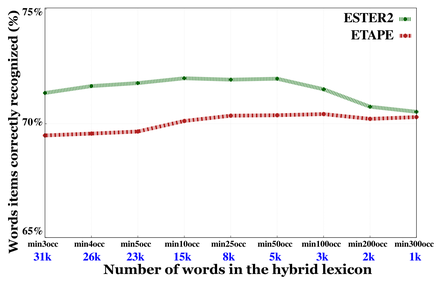


Retrieving the message carried out by the speech signal

How many words are generated by the decoder?

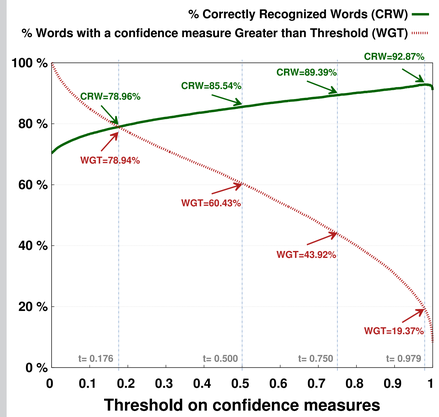


Among these words, how many of them were correctly recognized ?

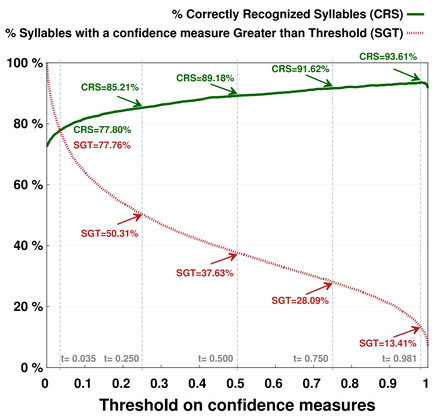


Can the confidence measures identify correct items?

correctly recognized words?



correctly recognized syllables?



Evaluation on ETAPE corpus, hybrid lexicon with 5k word entries

Conclusions

- the hybrid language model is a good compromise
- among the recognized words which have a confidence measure greater than 0.5, 85% are correctly recognized
- evaluations have also shown that the contribution of confidence measures on syllables is relevant only if there is a fairly significant amount of syllables in the language model

Future work

- investigate further confidence measures on the syllables units
- towards detection of error zones instead of item-based decision

RAPSODIE project, supported by the "Conseil Régional de Lorraine" (FEDER)

{firstName.lastName}@loria.fr