Comparison of approaches for an efficient phonetic decoding

Luiza Orosanu and Denis Jouvet

{luiza.orosanu, denis.jouvet}@loria.fr

Speech Group, LORIA
Inria, Villers-les-Nancy, F-54600, France

August 27, 2013

This study is part of the **RAPSODIE project** (http://erocca.com/rapsodie) and has received support from the "Conseil Régional de Lorraine" and from the "Région Lorraine" (FEDER)

Summary

Context

- Deafness
  - ∗ for **children**: can delay language development and cognitive skills
  - ∗ for **adults**: difficulty to find an employment, exercise and keep it
  - ∗ for **all**: social isolation

Context

- Deafness
  - for **children**: can delay language development and cognitive skills
  - for **adults**: difficulty to find an employment, exercise and keep it
  - for **all**: social isolation
- A speech recognition system adapted to deaf people's needs
  - improve communication between deaf people and their entourage
  - tool of socialization and/or integration in the workplace

Considerations

- Why consider a **portable solution** ?
  - ∗ could be used anywhere & anytime
  - ∗ could give real-time information to its owner

Considerations

- Why consider a **portable solution** ?
  - * could be used anywhere & anytime
  - * could give real-time information to its owner
- **Constraints** on considering an embedded device
  - * limited memory size
  - * limited computational power

Summary

Methodology

- **Objective**
  - ∗ find the best compromise between $\begin{cases} \text{computational cost} \\ \text{usability of results} \end{cases}$

Methodology

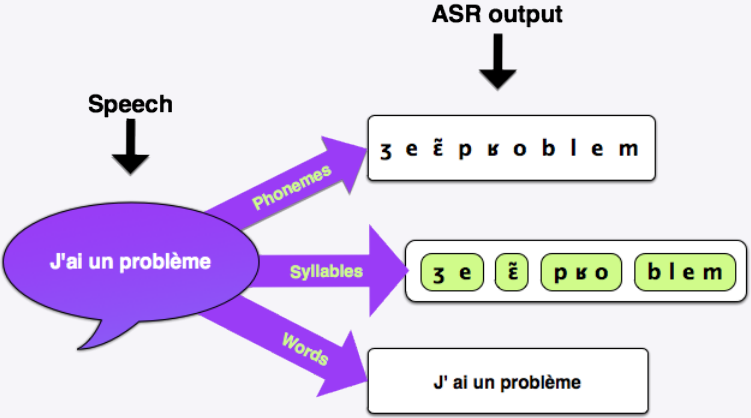
- **Objective**
  * find the best compromise between { computational cost / usability of results }

- **Approaches**
  * always use the **same acoustic units**
  * evaluate **3 different linguistic units**

    ⇒ different vocabularies & different language models

Acoustic unit	Linguistic unit
phoneme	phoneme
	syllable
	word

Comparison of linguistic units



Comparison of linguistic units

- phonemes
  - ∗ vocabulary : $<$ 40 phonemes for French
  - ∗ 3-gram language model : $<$ 1 MB

Lexicon entries

au $\Rightarrow$ au
b $\Rightarrow$ b
ge $\Rightarrow$ ge

Comparison of linguistic units

- phonemes
  - ∗ vocabulary : $<$ 40 phonemes for French
  - ∗ 3-gram language model : $<$ 1 MB

Lexicon entries
au $\Rightarrow$ au
b $\Rightarrow$ b
ge $\Rightarrow$ ge

- words
  - ∗ vocabulary : $\sim$ 97,000 words
  - ∗ 3-gram language model: $>$ 1 GB

absent $\Rightarrow$ a b s an
combiner $\Rightarrow$ k on b i n e
libre $\Rightarrow$ l i b r

Comparison of linguistic units

- phonemes
  - ∗ vocabulary : $<$ 40 phonemes for French
  - ∗ 3-gram language model : $<$ 1 MB
- **syllables**
  - ∗ vocabulary : $\sim$ 16,000 syllables
  - ∗ 3-gram language model : $<$ 10 MB
- words
  - ∗ vocabulary : $\sim$ 97,000 words
  - ∗ 3-gram language model: $>$ 1 GB

Lexicon entries

au ⇒ au
b ⇒ b
ge ⇒ ge

au_s ⇒ au s
b_l_au ⇒ b l au
o_r ⇒ o r

absent ⇒ a b s an
combiner ⇒ k on b i n e
libre ⇒ l i b r

Syllables

- Setup for **defining the syllables**
  * the training corpora is entirely **phonetized** (by forced alignment)
  * the sequence of phonemes is processed by the **syllabification tool**

Syllables

- Setup for **defining the syllables**
  - ∗ the training corpora is entirely **phonetized** (by forced alignment)
  - ∗ the sequence of phonemes is processed by the **syllabification tool**
- Rules of syllabification [Bigi et al,2010]
  - ∗ a syllable contains a single vowel (V)
  - ∗ a pause designates a syllable's boundary

[Bigi et al.,2010] Bigi, B., Meunier, C., Bertrand, R. and Nesterenko, I., "Annotation automatique en syllabes d'un dialogue oral spontané", Journées d'Étude de la Parole, 2010

Syllables

- Setup for **defining the syllables**
  - ∗ the training corpora is entirely **phonetized** (by forced alignment)
  - ∗ the sequence of phonemes is processed by the **syllabification tool**
- Rules of syllabification [Bigi et al,2010]
  - ∗ a syllable contains a single vowel (V)
  - ∗ a pause designates a syllable's boundary

Sequence of phonemes	Split position	Resulting syllables
VV	0	V V
VxV	0	V xV
VxxV	1	Vx xV
VxxxV	2	Vxx xV

[Bigi et al.,2010] Bigi, B., Meunier, C., Bertrand, R. and Nesterenko, I., "Annotation automatique en syllabes d'un dialogue oral spontané", Journées d'Étude de la Parole, 2010

Syllables

Example

ce qui s' est passé c' est que (...)

s k i s e p a s e s e k ← **forced alignment**

s_k_i **s_e** **p_a** **s_e** **s_e_k** ← **syllables**

Syllables

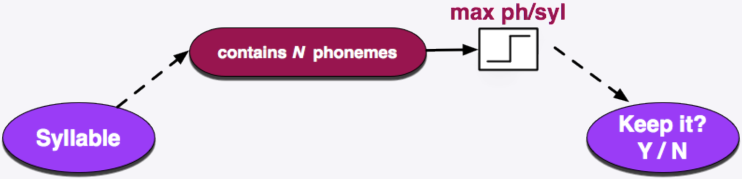
Example

ce qui	s'	est	passé		c' est que	(...)	
s k i	s e	p a	s e		s e k	← forced alignment	
s_k_i	s_e	p_a	s_e		s_e_k	← syllables	

⇒ The syllabification tool creates **syllables** and **pseudo-syllables**, which

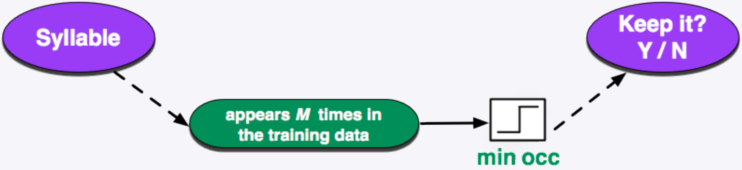
- ∗ take into account the **liaison & reduction** events
- ∗ are consistant throughout the entire training data

Syllables



- Reduce the number of (pseudo-)syllables by applying **two filters**
  - ∗ a **maximum number of phonemes** per syllable

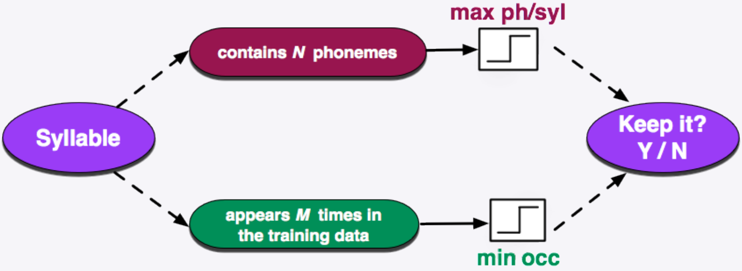
Syllables



- Reduce the number of (pseudo-)syllables by applying **two filters**

  * a **minimum number of occurrences** in the training data

Syllables



- Reduce the number of (pseudo-)syllables by applying **two filters**
  - ∗ a **maximum number of phonemes** per syllable
  - ∗ a **minimum number of occurrences** in the training data

⇒ create several different **lists of syllables**, by applying different thresholds for **each filter**

Summary

Experiments

- use a single type of **acoustic unit**
  - the phoneme
- use three different **linguistic units** (⇒ diffent vocabularies & LMs)
  - the phoneme
  - the syllable
  - the word
- test them on two French speech corpora
- study their phonetic decoding performance (PER)

LM = Language model

PER = Phonemes Error Rate

Data for Acoustic training

- **Train phonetic acoustic models**:
  - ∗ ESTER2 train set
  - ∗ ETAPE train set
  - ∗ EPAC train set

  $\Rightarrow$ 300h

Data for Acoustic training

- **Train phonetic acoustic models**:
  - ∗ ESTER2 train set
  - ∗ ETAPE train set
  - ∗ EPAC train set

  $\Rightarrow$ 300h

ESTER2 & EPAC
- ∗ French broadcast news, collected from radio channels
- ∗ prepared speech, plus interviews

ETAPE
- ∗ debates collected from various radio and TV channels
- ∗ spontaneous speech

Data for LM training

- **phoneme-based and syllable-based LM**

  → training from phonetic transcription

  - ∗ ESTER2 train set
  - ∗ ETAPE train set
  - ∗ EPAC train set

  ⇒ 12 million phonemes

  ⇒ 6 million syllables

Data for LM training

- **phoneme-based and syllable-based LM**

  → training from phonetic transcription

  - ∗ ESTER2 train set
  - ∗ ETAPE train set
  - ∗ EPAC train set

  ⇒ 12 million phonemes

  ⇒ 6 million syllables

- **word-based LM**

  → training from textual data

  - ∗ newspaper data
  - ∗ radio broadcast shows
  - ∗ French Gigaword corpus
  - ∗ web sources

  ⇒ more than 1.5 billion words

Data for Evaluation

- **Test** on:
  - ESTER2 development set (prepared speech) | $\Rightarrow$ 142,000 phonemes

Data for Evaluation

- **Test** on:
  - ESTER2 development set (prepared speech) ⇒ 142,000 phonemes
  - ETAPE development set (spontaneous speech) ⇒ 263,000 phonemes

Configuration

- SRILM tools
    - * build statistical Language Models

Configuration

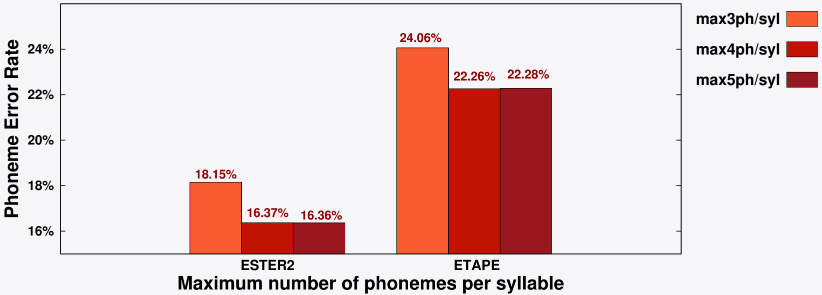
- SRILM tools
  - ∗ build statistical Language Models

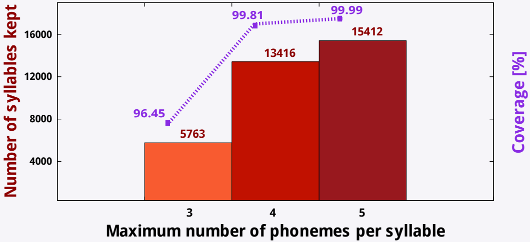
- MFCC acoustic analysis
  - ∗ compute 13 MFCC parameters per frame

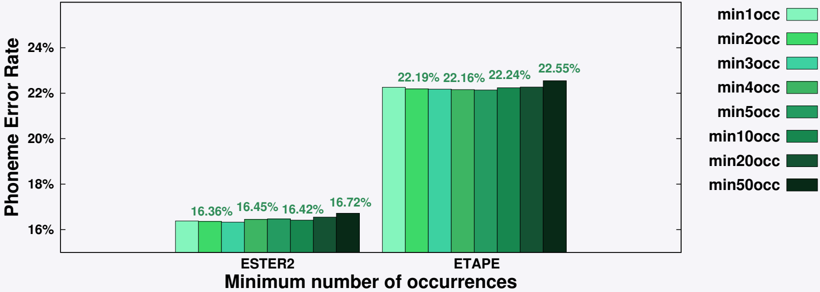
Configuration

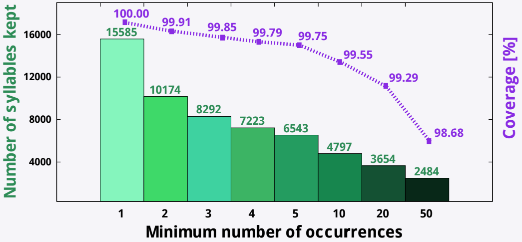
- SRILM tools
  - ∗ build statistical Language Models
- MFCC acoustic analysis
  - ∗ compute 13 MFCC parameters per frame
- Sphinx3 tools
  - ∗ train phonetic acoustic models
    - ⇒ Context dependent HMM acoustic models
      - 64 Gaussian mixtures
      - 7500 senones
      - adapted Male/Female
  - ∗ decode audio signals

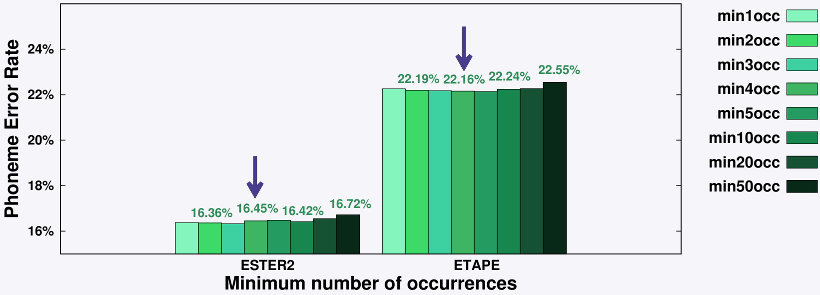
Results on the syllable-based LMs: filter by a maximum number of ph/syl

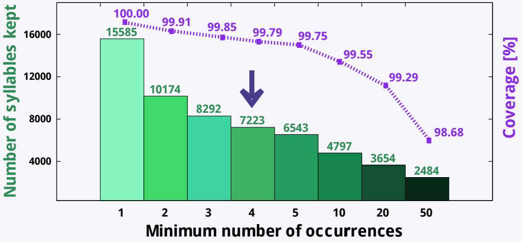




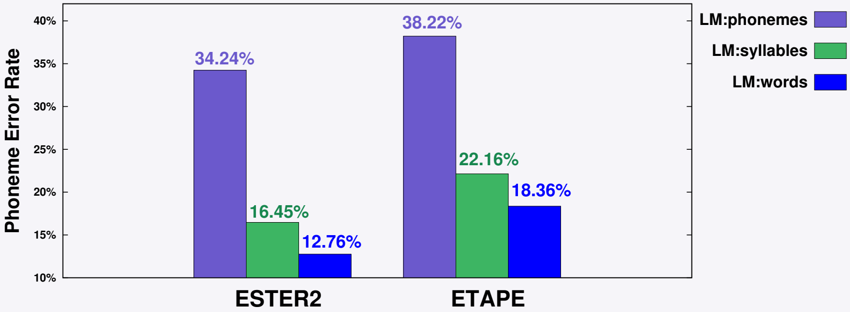

Phoneme Error Rate
24%
22%
20%
18%
16%
16.36%
16.45%
16.42%
16.72%
22.19%
22.16%
22.24%
22.55%
ESTER2
ETAPE
Minimum number of occurrences
min1occ
min2occ
min3occ
min4occ
min5occ
min10occ
min20occ
min50occ


Number of syllables kept
16000
12000
8000
4000
15585
10174
8292
7223
6543
4797
3654
2484
100.00
99.91
99.85
99.79
99.75
99.55
99.29
98.68
Coverage [%]
1
2
3
4
5
10
20
50
Minimum number of occurrences


Phoneme Error Rate
24%
22%
20%
18%
16%
16.36% 16.45% 16.42% 16.72%
22.19% 22.16% 22.24% 22.55%
ESTER2
ETAPE
Minimum number of occurrences
min1occc
min2occ
min3occ
min4occ
min5occ
min10occ
min20occ
min50occ


Number of syllables kept
16000
12000
8000
4000
100.00
99.91
99.85
99.79
99.75
99.55
99.29
98.68
15585
10174
8292
7223
6543
4797
3654
2484
1
2
3
4
5
10
20
50
Minimum number of occurrences
Coverage [%]

Overall results



ESTER2 : prepared speech

ETAPE : spontaneous speech

Summary

Conclusion

- phonetic n-gram language model
  - $\Rightarrow$ does not use much memory (< 1MB), nor computational power
  - $\Rightarrow$ does not give good results neither $\begin{cases} \sim 34\% \text{ PER ESTER2} \\ \sim 38\% \text{ PER ETAPE} \end{cases}$

LM = Language model

PER = Phonemes Error Rate

Conclusion

- phonetic n-gram language model
  - $\Rightarrow$ does not use much memory ($<$ 1MB), nor computational power
  - $\Rightarrow$ does not give good results neither $\begin{cases} \sim 34\% \text{ PER ESTER2} \\ \sim 38\% \text{ PER ETAPE} \end{cases}$

- word n-gram language model (LVCSR)
  - $\Rightarrow$ gives the best results $\begin{cases} \sim 12\% \text{ PER ESTER2} \\ \sim 18\% \text{ PER ETAPE} \end{cases}$
  - $\Rightarrow$ uses a lot of memory ($>$ 1GB) and computational power

---

LM = Language model

PER = Phonemes Error Rate

Conclusion

- phonetic n-gram language model
  - $\Rightarrow$ does not use much memory ($<$ 1MB), nor computational power
  - $\Rightarrow$ does not give good results neither $\begin{cases} \sim 34\% \text{ PER ESTER2} \\ \sim 38\% \text{ PER ETAPE} \end{cases}$
- **syllabic n-gram language models**
  - $\Rightarrow$ most frequent syllables $\rightarrow$ limited-size lexicon & LM ($<$ 10MB)
  - $\Rightarrow$ performance only 4% worse than the LVCSR $\begin{cases} \sim 16\% \text{ PER ESTER2} \\ \sim 22\% \text{ PER ETAPE} \end{cases}$
- word n-gram language model (LVCSR)
  - $\Rightarrow$ gives the best results $\begin{cases} \sim 12\% \text{ PER ESTER2} \\ \sim 18\% \text{ PER ETAPE} \end{cases}$
  - $\Rightarrow$ uses a lot of memory ($>$ 1GB) and computational power

---

LM = Language model

PER = Phonemes Error Rate

Future work

- find the best way of presenting the recognized information
  - ∗ phonemes
  - ∗ syllables
  - ∗ words or combinations

Thank you
for your attention !

Comparison of approaches for an efficient phonetic decoding

Luiza Orosanu and Denis Jouvet

{luiza.orosanu, denis.jouvet}@loria.fr

Speech Group, LORIA
Inria, Villers-les-Nancy, F-54600, France

August 27, 2013

This study is part of the **RAPSODIE project** (http://erocca.com/rapsodie) and has received support from the "Conseil Régional de Lorraine" and from the "Région Lorraine" (FEDER)