Towards automatic validation of pronunciation variants for speech recognition and synthesis

Colleen BEAUMARD, Nicolas PETITJEAN, Tom WYSOCKI
Supervisor: Denis JOUVET

Problematic

Speech synthesis and recognition have become important challenges for the future. They allow us to use voice commands on devices, but also to transcribe speech into text.

Objective: validation of word pronunciation variants by → alignment rules, inconclusive results, research of grapheme-to-phoneme conversion models to improve word prediction performance.

English pronunciation corpus

CMUdict - 134 304 words with their phonetic transcription (ARPABET symbols):

```
CATEGORIZE      K AE1 T AH0 G ER0 AY2 Z
CATEGORIZED     K AE1 T AH0 G ER0 AY2 Z D
CATEGORIZES     K AE1 T AH0 G ER0 AY2 Z IH0 Z
CATEGORIZING    K AE1 T AH0 G ER0 AY2 Z IH0 NG
CATEGORY        K AE1 T AH0 G AO2 R IY0
CATELLI         K AH0 T EH1 L IY0
CATENA          K AH0 T IY1 N AH0
CATER           K EY1 T ER0
CATERED         K EY1 T ER0 D
CATERER         K EY1 T ER0 ER0
```

0 : no stress - 1 : primary stress - 2 : secondary stress

French pronunciation corpus

BDLex - 337 550 words with their phonetic transcription (in SAMPA):

Right side		Upside down	
ballottement	b a l O/ t @ m a~	tnemettollab	a~ m @ t O/ l a b
ballottements	b a l O/ t @ m a~	stnemettollab	a~ m @ t O/ l a b
balsa	b a l z a	aslab	a z l a b
balsas	b a l z a	saslab	a z l a b
bambochade	b a~ b O/ S a d	edahcobmab	d a S O/ b a~ b
bambochades	b a~ b O/ S a d	sedahcobmab	d a S O/ b a~ b
bambocheur	b a~ b O/ S 9 R	ruehcobmab	R 9 S O/ b a~ b
bambocheurs	b a~ b O/ S 9 R	sruehcobmab	R 9 S O/ b a~ b
bambocheuse	b a~ b O/ S 2 z	esuehcobmab	z 2 S O/ b a~ b
bambocheuses	b a~ b O/ S 2 z	sesuehcobmab	z 2 S O/ b a~ b

Processing of the corpus

Train 70%	Development 15%	Test 15%

For the English lexicon, *test* (with stress) and *test bis* (without stress)

For the French lexicon, each file in right side and upside down versions.

For the models of the **neural network** approach and the **statistical** approach, we evaluate the prediction performances of the files with 2 scores:
WER = *Word Error Rate*, the error rate per word and **PER** = *Phoneme Error Rate*, the error rate per phoneme.

Alignment rule approach

Alignment of graphemes and phonemes using predefined rules. If insufficient, add new rules:

um → o~ columbarium → k O/ l o~ b a R j O m
z → ts breitschwanz → b R a j t S v a t s
ue → ju fuel → f j u l

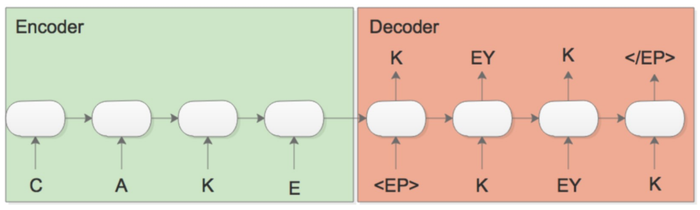
Résultats

Do not detect mispronunciations
→ even if incorrect predictions, alignment possible:

abstinent a b s t i n

Neural network approach

It is given a sequence of letters as input and the network provides a sequence of phonemes as output.



Software : OpenNMT

- Free software for neural networks;
- Use of a YAML configuration file (where to find the files needed for training).

Statistical approach

Statistical method for predicting a phoneme from a sequence of letters.

"mixing" [mɪksɪŋ] = m [m] | i [ɪ] | x [ks] | i [ɪ] | n [ŋ] | g —

Software : Sequitur

- Each element is linked to a sequence of letters and a sequence of phonemes, sequences that allow the reconstruction of the word with its pronunciation → Sequence Joined;
- Based on a grapheme-phoneme alignment.

English and French lexicon results

English Lexicon

Modele	PER	WER
With consideration of stress	10.93%	41.58%
Without taking stress into consideration	8.71%	36.86%

French Lexicon

Modele	PER	WER
OpenNMT right side	0.59%	3.05%
OpenNMT upside down	0.76%	4.24%
Sequitur right side	0.53%	3.15%
Sequitur upside dow	0.50%	2.92%

Models combination results (French lexicon)

Combination of the models

Number of identical predictions among the models	Number of predictions out of 51612	Proportion of configurations of a total in percent
4	48 165	93.32%
3-1	2098	4.06%
2-2	917	1.78%
2-1-1	424	0.82%
1-1-1-1	8	0.02%

Examples:

3 - 1	: a b l y t j o~	a b l y s j o~	a b l y s j o~	a b l y s j o~
2 - 1 - 1	: a~ m 9 l	a m 9 l	a~ m 2 l	a~ m 2 l

Conclusion

- Rules not sufficient to validate pronunciation variants;
- Neural network and statistical models perform well, slightly better when combined: (**PER** : 0.47% and **WER** : 2.75%).