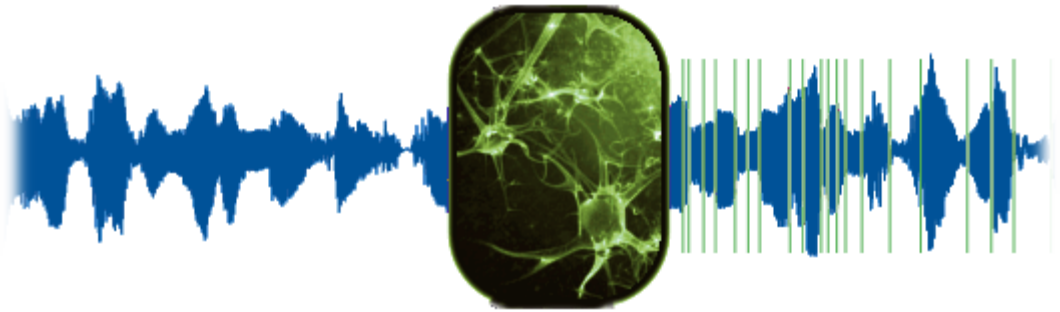
CNN-based phone segmentation experiments in a less-represented language

Céline Manenti, Thomas Pellegrini, Julien Pinquier
Université de Toulouse; UPS; IRIT; Toulouse, France



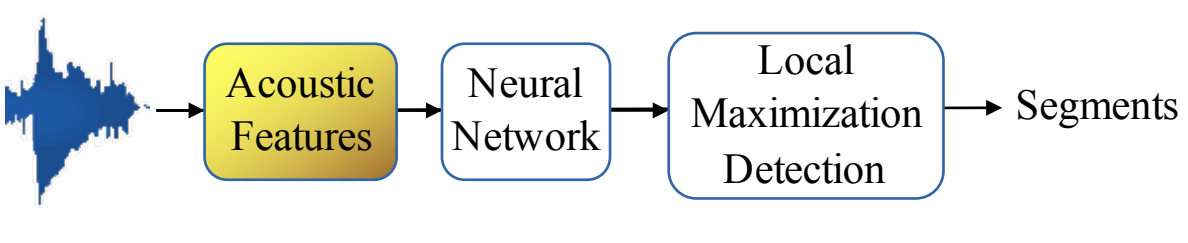
Introduction

- Why segmenting a speech signal into phone-like units?
  - Help the tedious and costly task of manual phonetic transcription (*e.g.* BUCKEYE corpus annotation took more than 2 years)
  - Discover phone-like units in an unsupervised fashion
  - Help for the development of ASR systems for low-resourced languages

Introduction

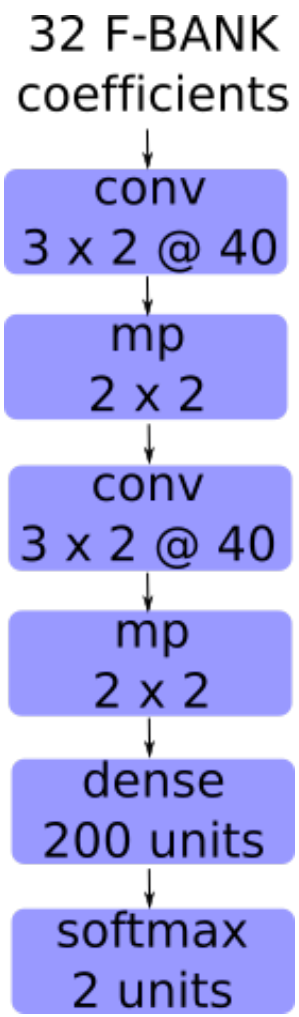
- Why neural networks?
  - Capability to solve many tasks (*e.g.*, NN mimic filter bank extraction [Bhargava et *al*, 2015])
  - State-of-the-art technique in ASR and in many speech processing tasks (*e.g.*, beats GMM in Vowel mispronunciation detection [Joshi and *al*, 2015])
- In this work: Convolutional Neural Networks

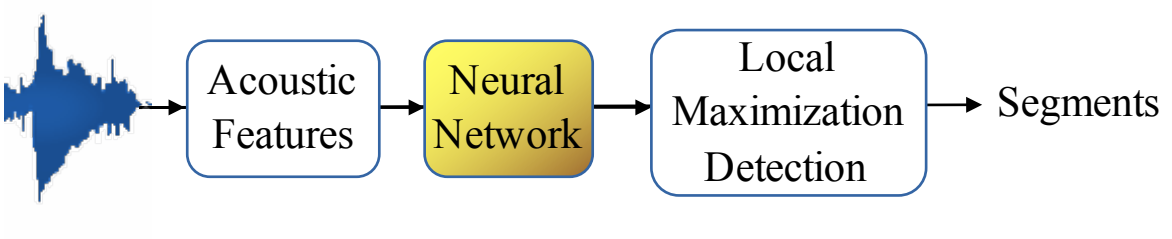
System Description



- F-BANK: 32 coefficients
- Window size: 16ms
- Hop size: 4ms
- Context:18 neighboring frames (88ms)

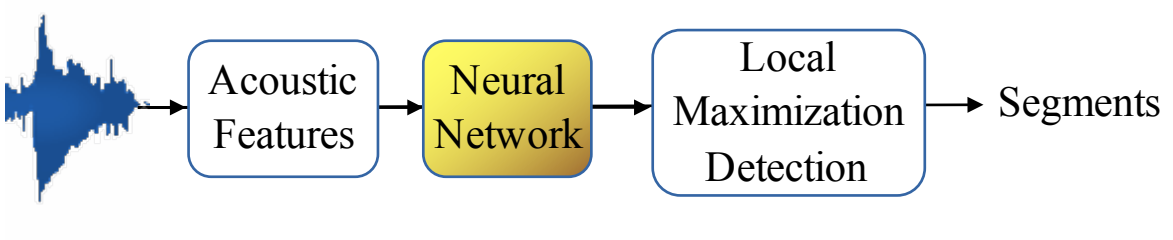
System Description



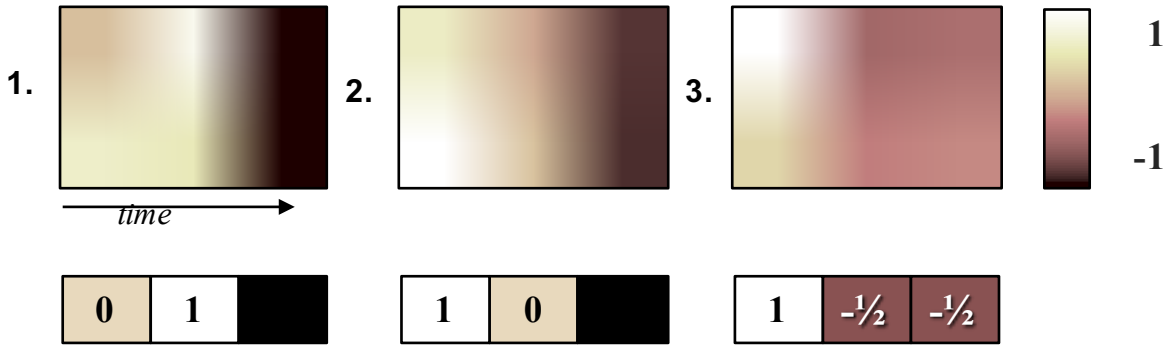


- Tests with 15 to 120 filters and 50 to 300 neurons
- Only 1% gain with 120 filters and 300 neurons

System Description



- The first CNN layer approximates derivation computations



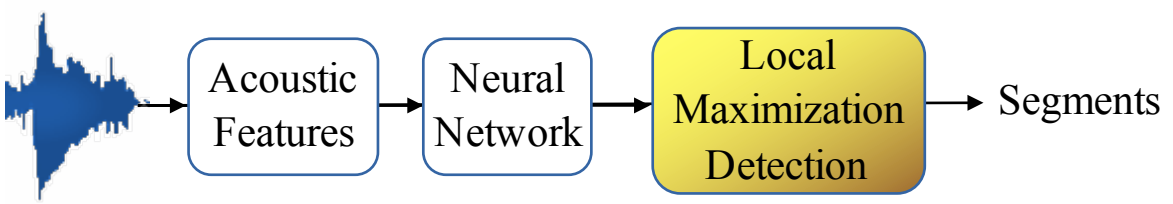
1. $s(t) - s(t+1)$

2. $s(t-1) - s(t+1)$

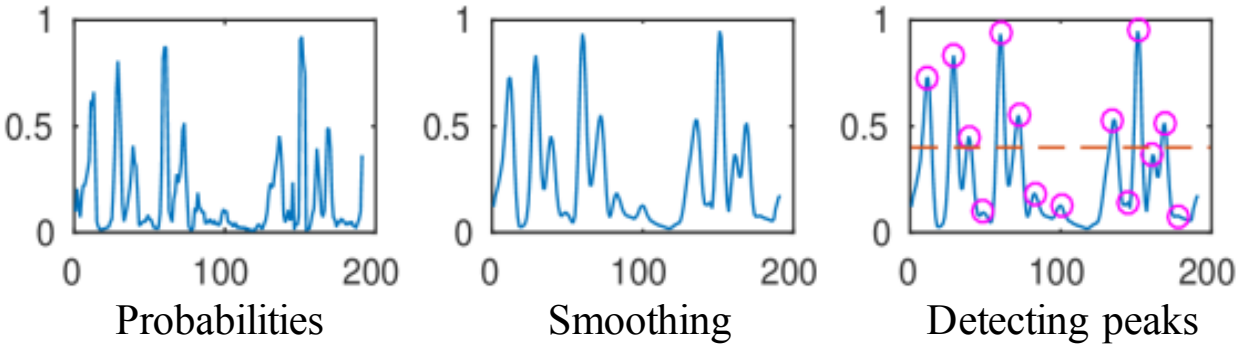
3. $s(t-1) - \frac{1}{2}\, s(t) - \frac{1}{2}\, s(t+1)$

- Model output: frame-level probability

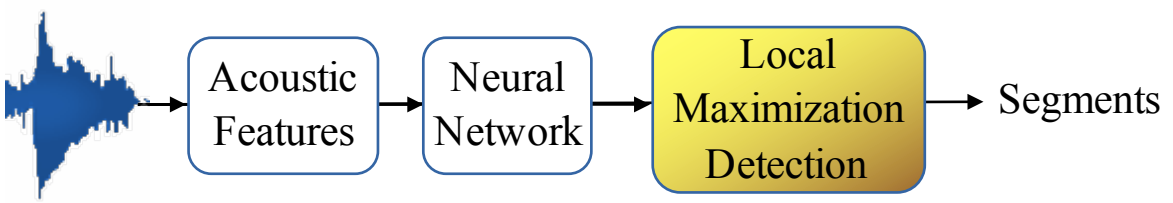
System Description



- Basic probability threshold: 0.5
- Problem: several boundaries may be predicted for neighboring frames
- Solution: select the frame with the maximum probability locally: need to retrieve peaks

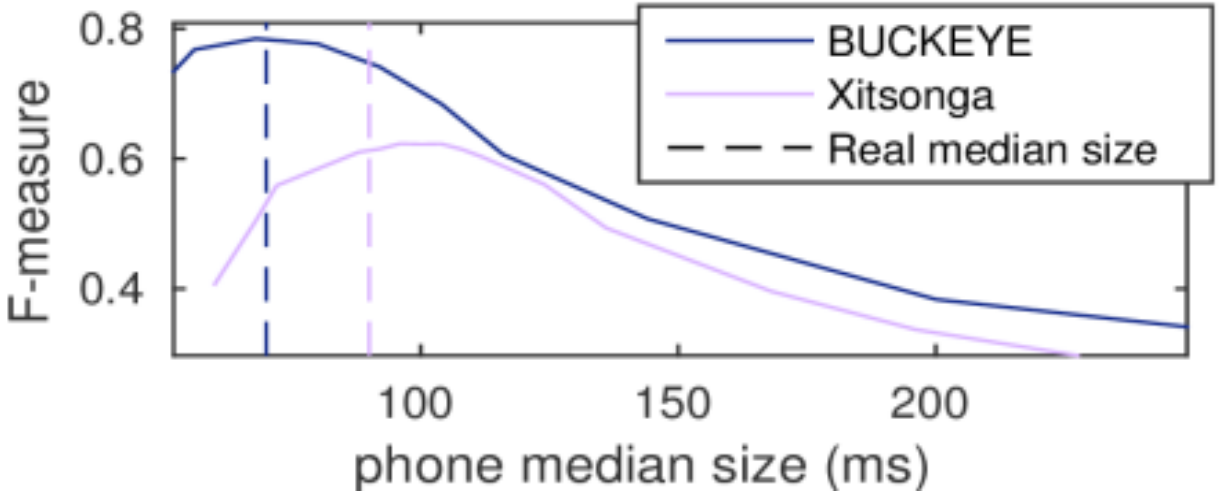


System Description



- Problem: unbalanced number of positive and negative samples
- Solution: threshold value tuning

  ⬅➡ choosing a mean segment duration



Experiments

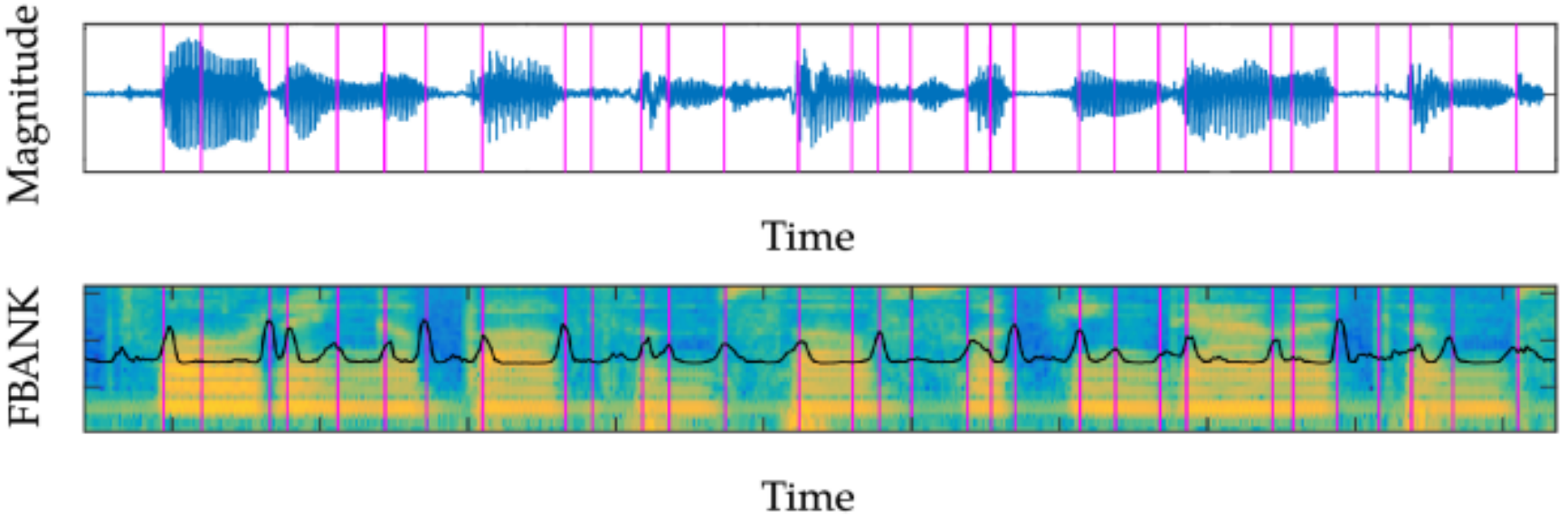
BUCKEYE

- American English, spontaneous speech, recorded in studio
- 40 speakers, ½ hour each
  - Train: 10 hours, 22 speakers
  - Development: 3 hours, 6 speakers
  - Test: 5 hours, 12 speakers

Xitsonga

- Low-resourced language, read speech, recorded on smartphones
- 4 speakers, ½ hour in total
  - Train: 20 minutes
  - Test: 10 minutes

Experiments

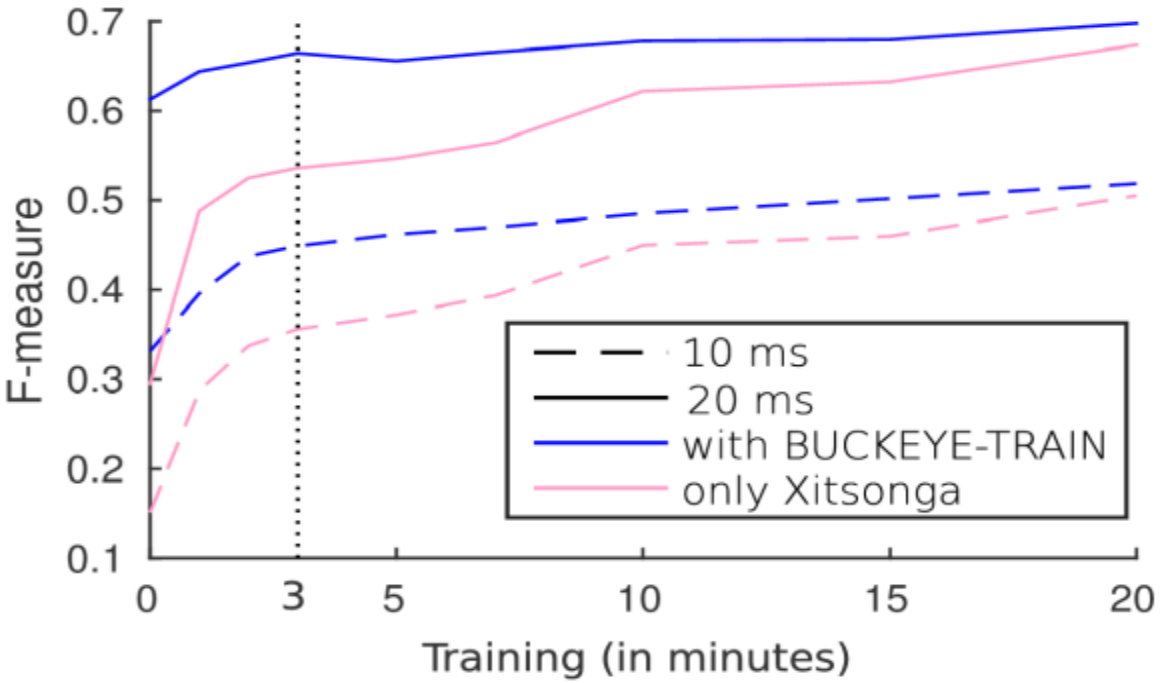


Experiments

	Annotators	CNN
10ms	0.62	0.68
20ms	0.79	0.79

- Performance similar to the inter-agreement rate between human annotators
- High precision, low recall

Experiments



- Impact of:
  - using a model trained on BUCKEYE (*vs*. new model)
  - Xitsonga training data size
  - Tolerance margin

Conclusions

- CNN well adapted to phone segmentation
- Performance similar to the inter-agreement rate
- Model portability: good results obtained on a low-resourced language using a model trained with English data
- Ongoing work:
  - In-depth analysis of the segmentation errors
  - Segmentation used in unsupervised discovery of acoustic units

Thank you!

Q&A

celine.manenti@irit.fr