Simple neural representations of speech for voice activity detection and speaker tracking in noisy archives

David Doukhan and Jean Carrive - {ddoukhan, jcarrive}@ina.fr
French National Institute of Audiovisual – http://www.ina.fr
4th International Conference on Statistical Language and Speech Processing - SLSP 2016

VAD and Speaker Tracking

Voice Activity Detection: prerequisite to speech analysis

Speaker Tracking: infer which portions of a recording correspond to a known speaker

Goals in an archiving workflow

Indexation

Enhanced Media

Speech time count

Limitations

Temporal resolution

Noise management

Dependent on the recording conditions

Material: Rivonia Trial (1963-1964)

Dictabelt Recordings

Floppy vinyl cylinders

30 minute recordings

Digitized using Henri Chamoux's Archeophone

Medium-specific timbral properties

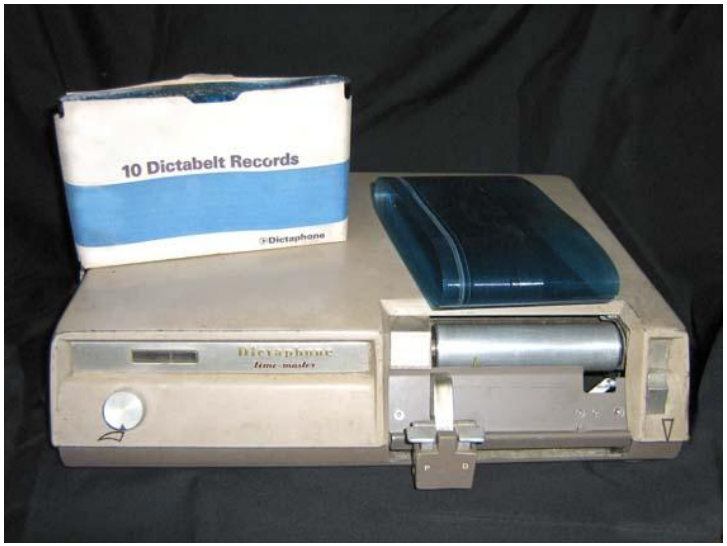
Digitization artifacts

Omni directional microphone

Noise

Small Speech turns

Speech superposition



Models adapted to this medium should be built!

VAD and Speaker tracking algorithm

1. Feature extraction: spectrum power (FFT), mel-frequency cepstrum (MFC) or Mel-frequency cepstral coefficients (MFCC)
2. Normalization: Zero Component Analysis (ZCA) whitening
3. Neural Space Projection: 1 layer convolutional neural network, with RELU activations, trained with K-Means (unsupervised)
4. Frame concatenation: frame context management
5. Probabilistic support vector machine (SVM) classification
6. Post-processing: Viterbi decoding of SVM output

Steps 2 and 3 inspired by

Coates, A., Lee, H., & Ng, A. Y. (2010). An analysis of single-layer networks in unsupervised feature learning. Ann Arbor, 1001(48109), 2.

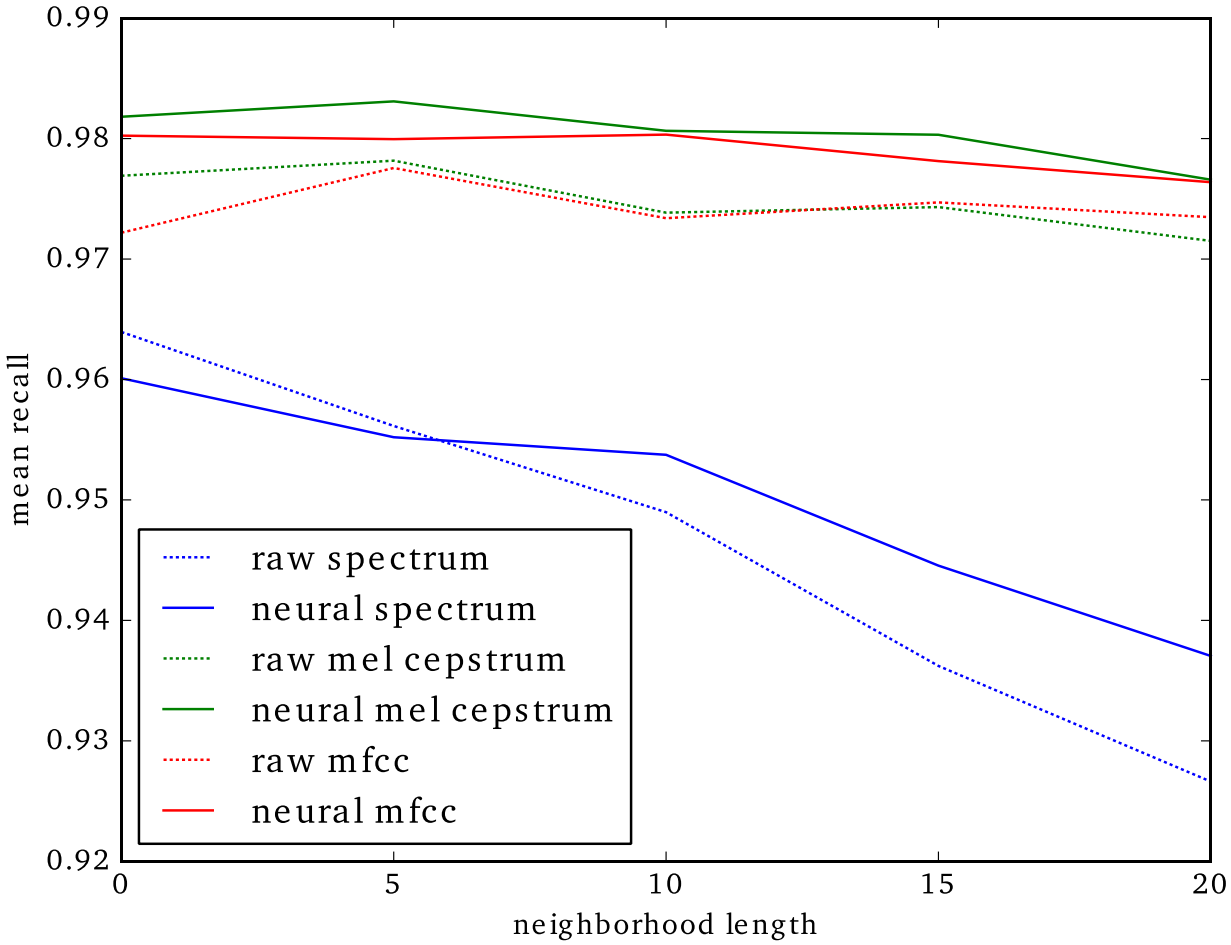
Evaluation Corpus

2 recordings
2 speakers per recording
1 minute of annotated speech per speaker
80 seconds of non speech per recording

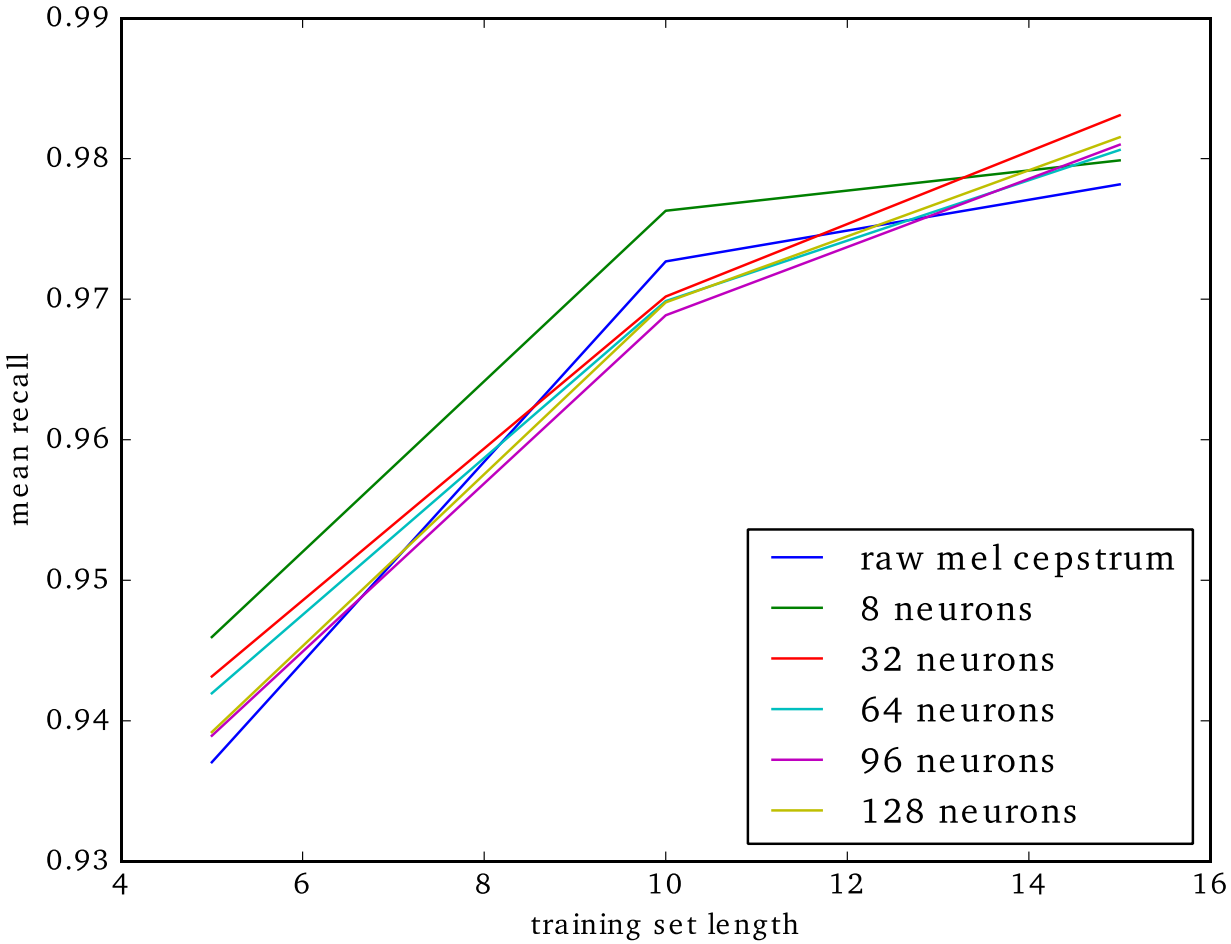
Evaluation Protocol

3 frame-level features (spectrum, MFC, MFCC)
Training set length: 5, 10 and 15 seconds
Raw features versus neural representations (8, 16, 32, 64, 96, 128)
Neighborhood context: +- 0, 5, 10, 15, 20
Evaluation Metric: Mean Recall

Voice Activity Detection Evaluation



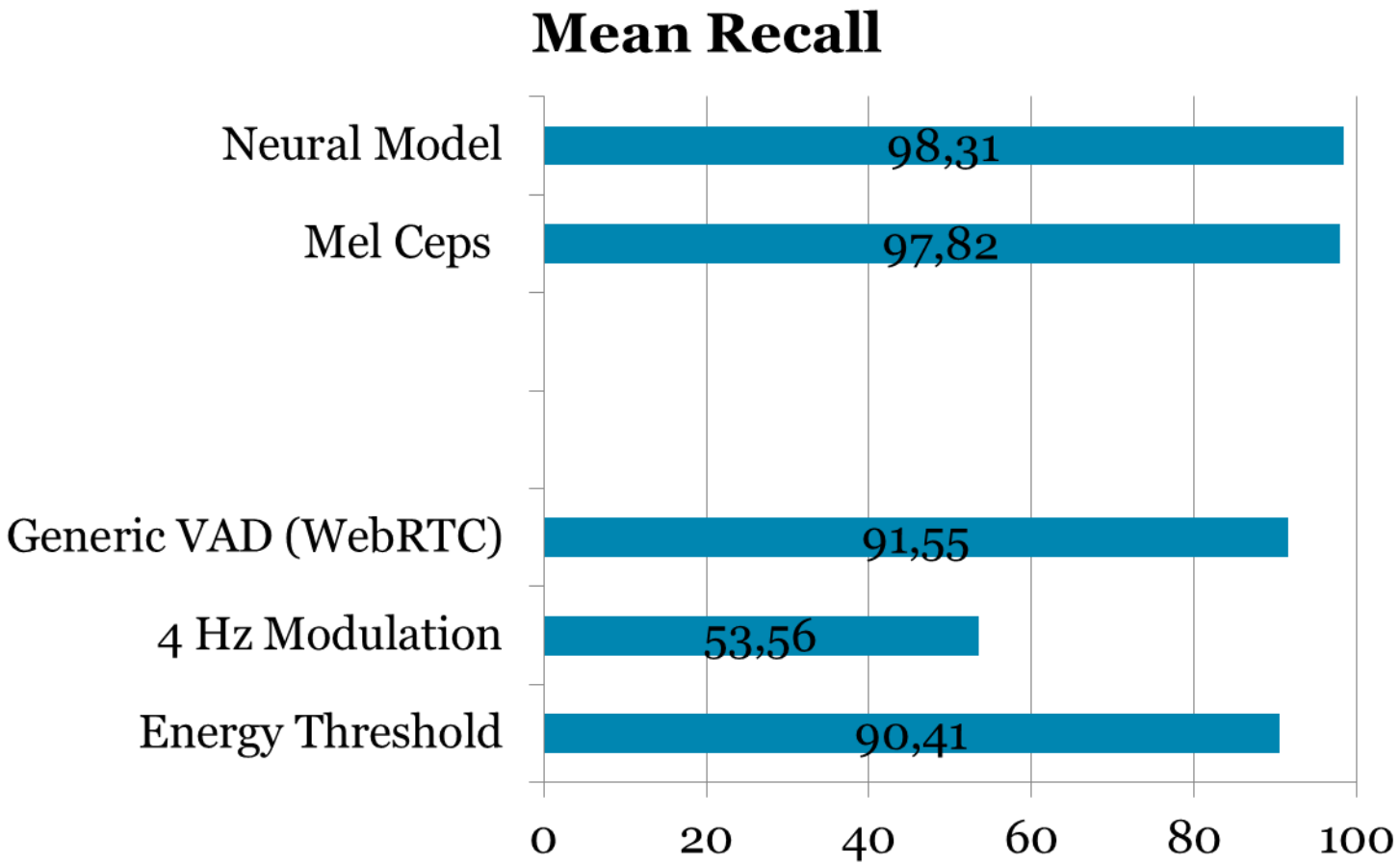
VAD training set size impact



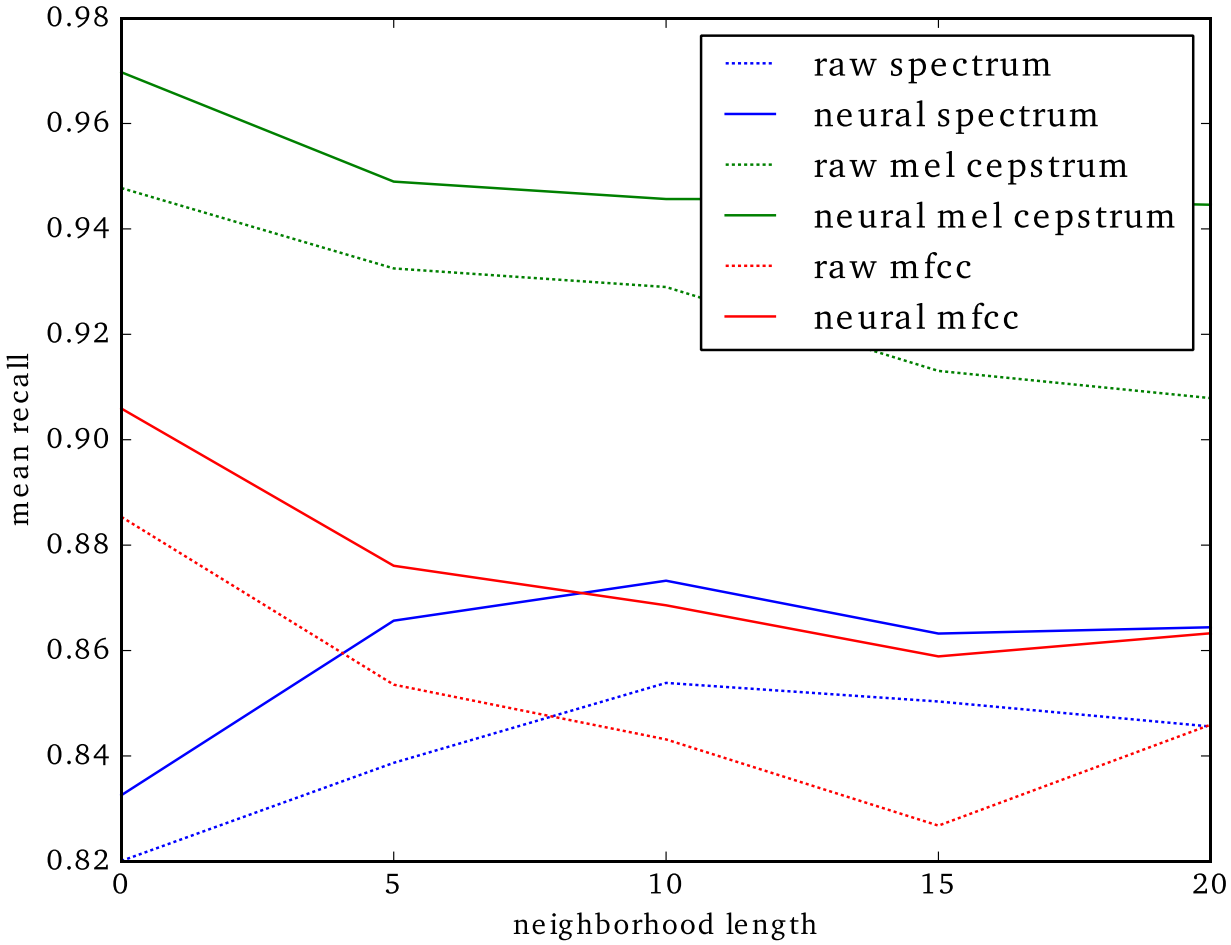
Voice Activity Detection system benchmark

Proposed Approach

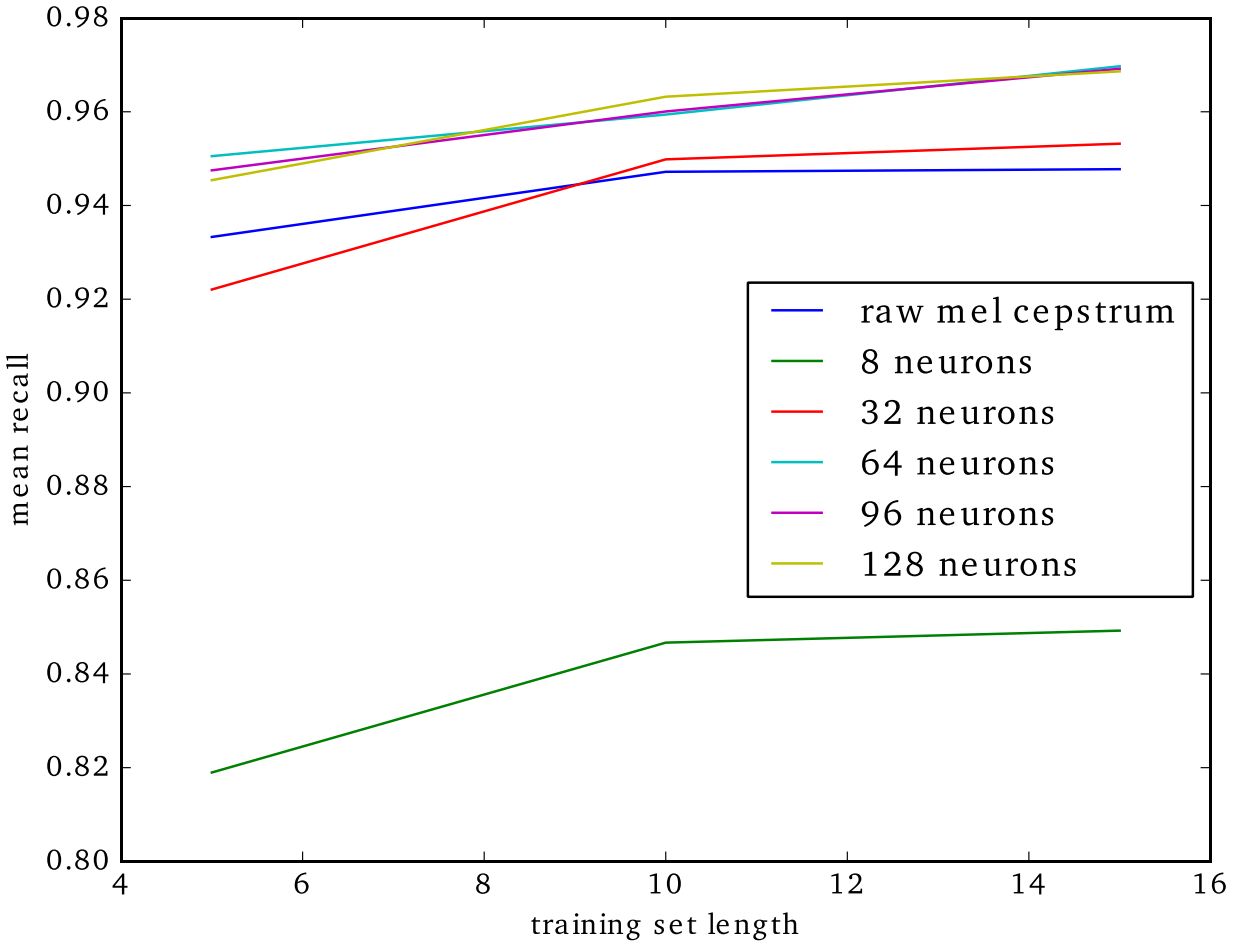
Baseline Approaches + Viterbi decoding



Speaker Tracking evaluation



Speaker tracking training set size impact



Conclusion & Future Work

Models fitted to the recording medium were proposed

Neural representation of speech, obtained through unsupervised procedure, allowed to obtain better performances:

+0,49 Mean recall on VAD

+2,2 Mean Recall on Speaker Tracking

Future Work

Investigate the use of larger training sets

Evaluate this approach on clean speech

Use neural features for unsupervised speech turn structuration (diarisation)

Investigate deeper architectures