Acoustic modeling of speech waveform based on multi-resolution, neural network signal processing

Zoltán Tüske, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition Group,
RWTH Aachen University, Germany

Outline

Introduction

Towards multi-resolution NN signal processing

Experimental Setup

Experimental Results

Weight analysis

Conclusions

Introduction

Before the recent advance of deep neural network in acoustic modeling (AM):

- Manually designed feature extraction methods are based on:
  - Physiology, [von Békésy, 1960], psychoacoustics [Fletcher and Munson, 1933], trial-and-error [Furui, 1981]
- MFCC [Davis and Mermelstein, 1980], PLP [Hermansky, 1990], GT [Schlüter et al., 2007].

Current trend in neural network based AM:

- Learn the complete feature extraction from data, as part of the AM.
  - Single channel: [Palaz et al., 2013, Tüske et al., 2014] [Golik et al., 2015, Zhu et al., 2016, Ghahremani et al., 2016].
  - Multi-channel, incl. beamforming: [Hoshen et al., 2015, Li et al., 2016].

- Usually: efficient modeling of direct waveform needs large amount of data.

State-of-the-art direct waveform AM

Similar to standard features:

- Starts with time-freq. (TF) decomposition by 1-D convolution, like STFT or Gammatone filters.

$$y_{k,t} = \sum_{\tau=0}^{N_{\text{TF}}-1} s_{t+\tau-N_{\text{TF}}+1} \cdot h_{k,\tau} \tag{1}$$

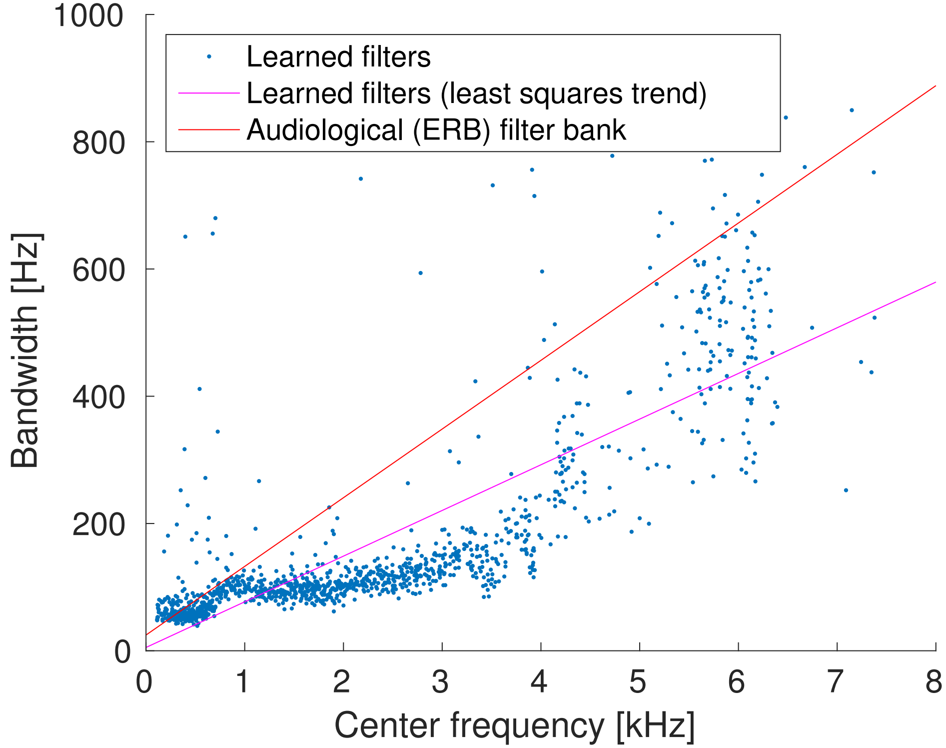
  - s_t: input signal, sampled at 16kHz.
  - $y_{k,t}$: optionally sub-sampled filter-output.
  - $h_{k,t}$: mirrored FIR filter impulse response, $N_{\text{TF}} = 512 = 32ms@16kHz$.

- Followed by envelope extraction
  - Rectification, low-pass filtering, and sub-sampling:
    - Non-parametric: max [Hoshen et al., 2015], average [Sainath et al., 2015], p-norm [Ghahremani et al., 2016] pooling.
    - Non-overlapping stride: sub-sampling at a single fixed $\sim$10ms rate.

Introduction

Issue:

- Learned TF filters have varying bandwidth
- Estimated bandwidth vs. center frequency [Tüske et al., 2014]:



- Fix rate subsampling might lead to under-sampling of broader band-pass filters, non-recoverable.

Introduction

In this study:

- Generalizing the envelop extractor/down-sampling block.
  - Making it trainable.
  - See also network-in-network approach of [Ghahremani et al., 2016]
- Allowing the network to learn multi-resolution spectral representation.
  - See also multi-scale max-pooling approach of [Zhu et al., 2016].

Towards multi-resolution NN signal processing

Parametrized envelope extraction:

- By trainable FIR low-pass filters.

$$x_{i,k,t} \stackrel{\text{FIR}}{=} f_2\left(\sum_{\tau=0}^{N_{\text{ENV}}-1} f_1\left(y_{k,t+\Delta t_{TF}\cdot\tau-N_{\text{ENV}}+1}\right)\cdot l_{i,\tau}\right) \tag{2}$$

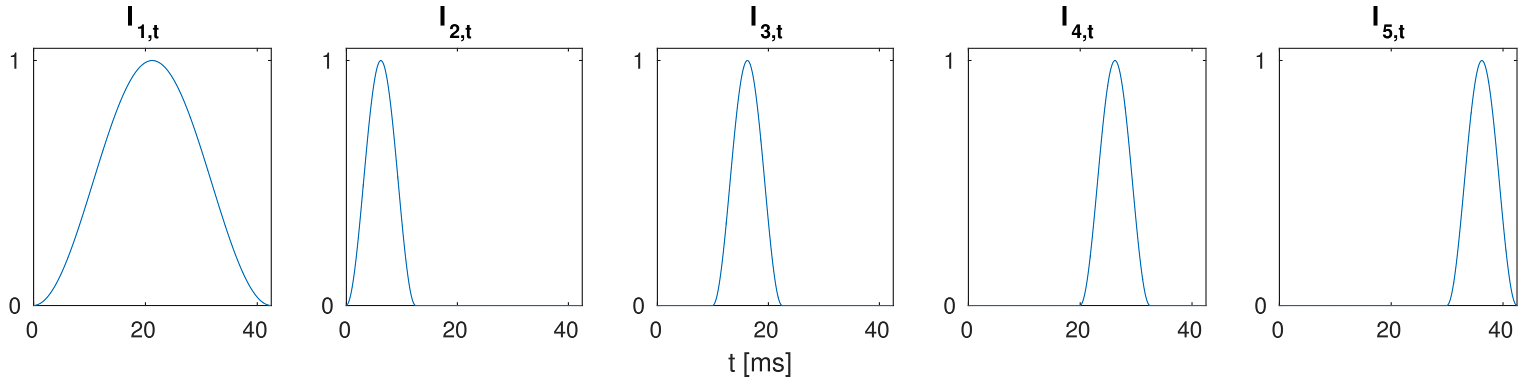
  - $f_1(y_{k,t})$: rectified TF filter output subsampled at $\Delta t_{TF} = 10 = 0.625ms@16khz$ step, (contains very fine time structure, fits for TF filter with up to 800Hz bandwidth)
  - f_2: incorporates additional signal processing steps, e.g. root or logarithmic compression.
  - $l_{i,t}$: trainable low-pass filter, $N_{\text{ENV}} = 16..160$, up to 100ms (long).
  - $x_{i,k,t}$ evaluated at $\Delta t_{ENV} = 16\cdot 10$, $10ms@16kHz$ rate.

- 2^{nd} level of 1-D convolution.
- Parameters are shared in time and between the TF filters.
- Although output sampled at fixed 10ms rate, the structure allows multi-resolution processing.

Towards multi-resolution NN signal processing

The proposed structure allows:

- The learning of multi-resolution processing of critical bands, e.g.:
  - E.g.: assuming 5 envelope filters, $i = 1..5$.
  - Access to both fast and low rate sampled critical band.
  - Localization, shifting the ,,faster" low-pass filter within the analysis window.

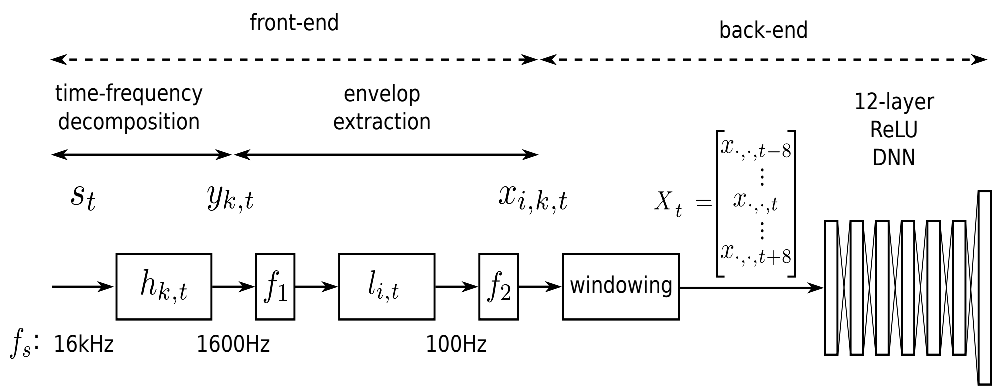


- Wavelet-like processing:
  - Exhaustive combination of envelope processing and TF filters, non-orthonormal basis.
  - Orthonormal sub-space can be selected from $x_{i,k,t}$.
  - We let the NN decide which elements of $x_{i,k,t}$ contain useful information.

Experimental Setup

- Models evaluated on an English broadcast news and conversation ASR task, reporting WER.
- Training data consisted of 250 hours of speech, 10% selected for cross-validation.
- Dev. and eval sets contain 3 hours of speech.

- Back-end (BE): a hybrid 12-layer feed-forward ReLU MLP, 2000 nodes per layer.
  - 17-frame window.
  - 512-dim. low-rank factorized first layer.
  - Dimension of X_t is up to 150x20x17 = 51000.



- Models are trained using:
  - Cross-entropy, SGD, momentum, $L2$, and discriminative pre-training.

Experimental Results

Comparison of envelope filter types

- 50 TF filters, single envelope filter.
- $f_1(.) = Abs(.)$, $f_2(.) = \sqrt[2.5]{Abs(.)}$

$I_{i,t}$ type	N_{ENV}	WER dev	WER eval
max	16	14.4	19.9
	25	14.3	19.8
	40	14.4	19.7
FIR	40	14.1	19.8
Gammatone		13.5	18.4
time-signal DNN		15.1	20.5

- Overlapping ($N_{\text{ENV}} > 16$) max pooling performs slightly better.
- Trainable element is as effective as max pooling.
- More (+100) TF filters lead to further modest improvement: 0.4% on eval set.

Experimental Results

Effect of envelope detector ($l_{i,t}$) size, and non-linearities:

#env. filters ($l_{i,t}$)	N_{ENV} sample	N_{ENV} ms	f_1	f_2	#param*	WER dev	WER eval
5	40	25	Abs(.)	-	7.5M	14.2	19.6
			Abs(.)	Abs(.)		14.2	19.3
			Abs(.)	$\sqrt[2.5]{Abs(.)}$		13.7	18.7
			$\sqrt[2.5]{Abs(.)}$	Abs(.)		13.8	18.7
10	80	50	Abs(.)	Abs(.)	14M	13.9	19.0
				$\sqrt[2.5]{Abs(.)}$		13.9	19.0
20	160	100	Abs(.)	Abs(.)	27M	14.3	19.3
				$\sqrt[2.5]{Abs(.)}$		14.4	19.6
Gammatone					1.7M	13.5	18.4

*up to 1st back-end layer

- Using multiple envelope filters is closing the WER gap to Gammatone.
- The root compression seems to be important only if $N_{\text{ENV}} < 10$.

Experimental Results

Effect of the segment-wise mean-and-variance normalization:

- Freezing the front-end, and retraining the back-end model on the normalized features.

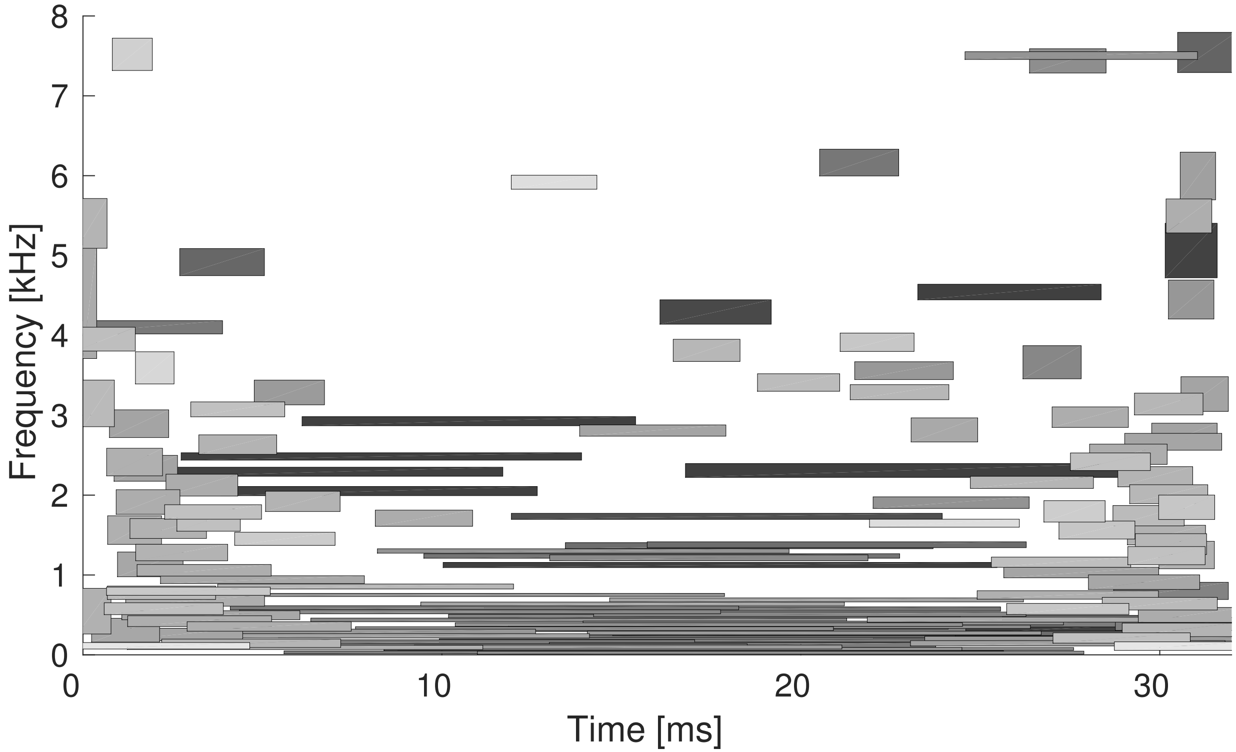
front-end		normalization		WER [%]	
type	dim.	mean	variance	dev	eval
NN	512			13.7	18.7
		×		13.7	18.6
		×	×	13.5	18.5
GT	70x17			13.5	18.4
		×		13.1	17.8
		×	×	13.2	17.9

- Segment level normalization improves NN front-end, but less effective than with Gammatone.
- Increased performance gap between the Gammatone (GT) and direct waveform models.

Weight analysis

Analyzing the time-frequency decomposition layer ($h_{k,t}$).

- Plotting time-frequency patches in the 32ms analysis window (operates at 0.625ms shift).
- Estimating center freq., pulse-, and bandwidth for each (150) band-pass.
- The grayscale intensity is proportional to patch surface.

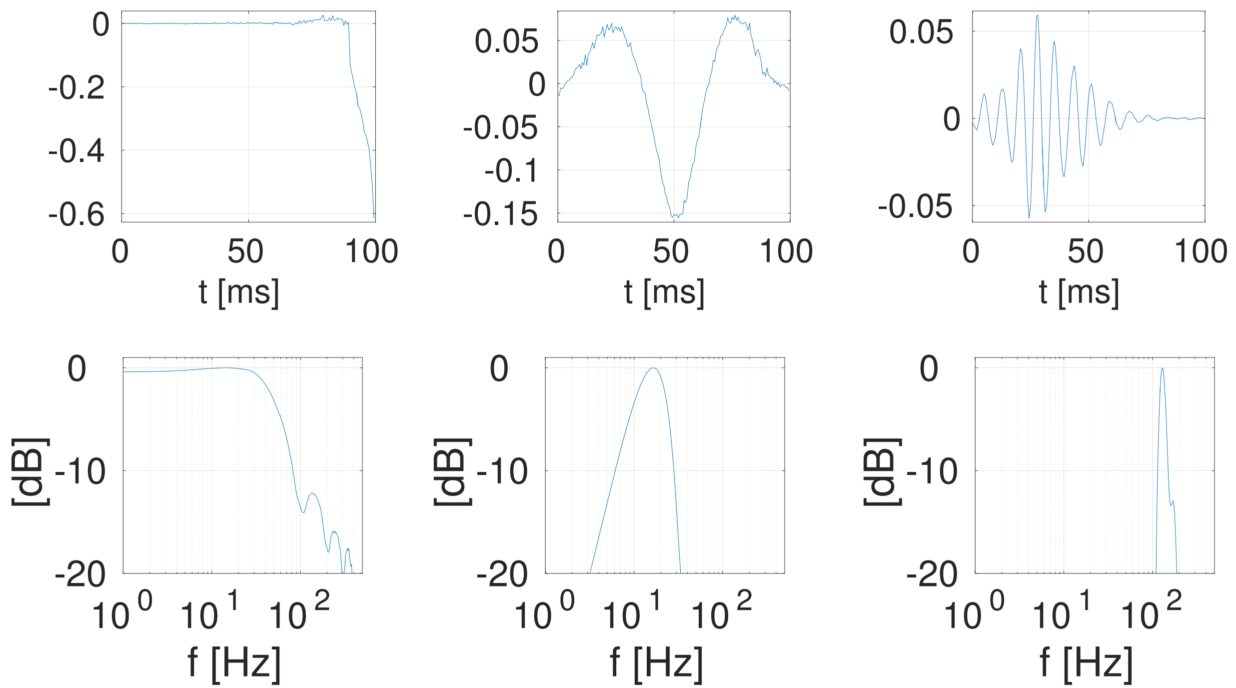


- Multi-resolution: each frequency band is covered by various band-pass filters.

Weight analysis

Analyzing the envelope extractor layer ($l_{i,t}$):

- Examples of $l_{i,t}$ and below its Bode magnitude plot:

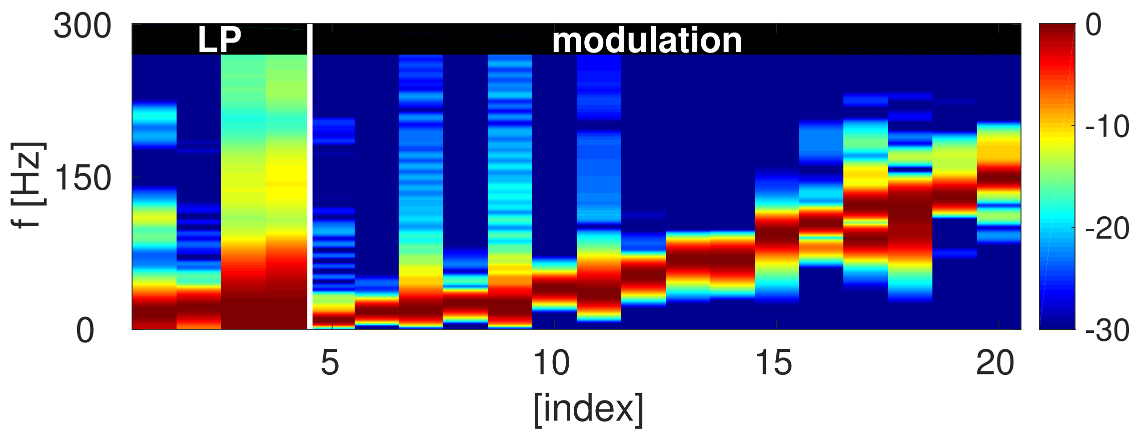


- Surprisingly, besides low-pass also many band-pass filters: modulation spectrum.

Weight analysis

Analyzing the envelope extractor layer:

- $l_{i,t}$ can be split to low-pass (LP) and modulation filters.
- Filters can be sorted by the cutoff or center frequencies.
- Plotting amplitude spectrum of the reordered $l_{i,t}$.

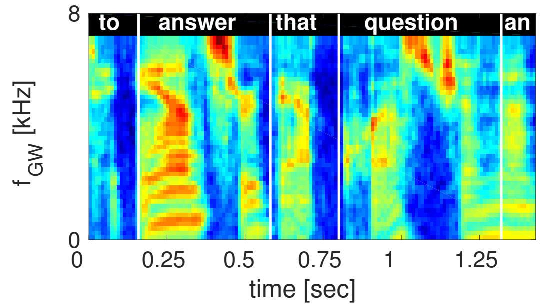


- Multiple low-pass filter, according to variable bandwidth of TF filters.
- Modulation filter frequencies are clearly below 150Hz.
  - Research studies on modulation spectrum suggest only 20-40Hz.

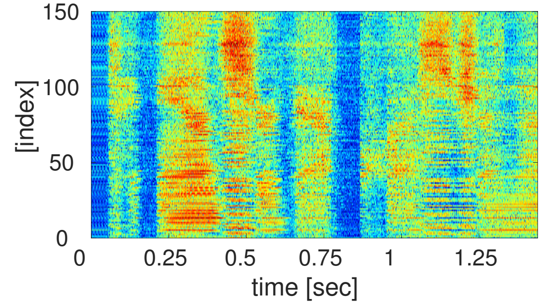
Weight analysis

Comparison of standard Gammatone and NN spectrograms (CRBE):

GT CRBE



- resolution: 10ms

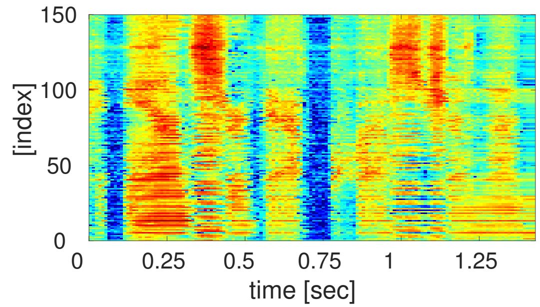


NN CRBE

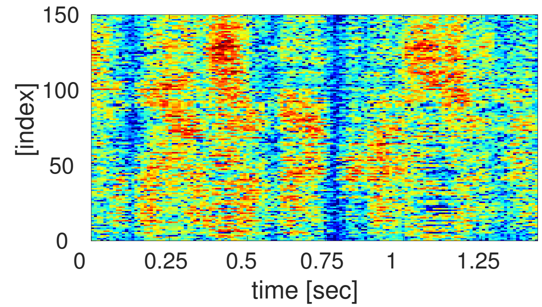
- $f_1(y_{k,t})$ resolution: 0.625ms

NN spectrograms after low-pass (LP) and modulation (MOD) filtering ($x_{i,k,t}$):

NN LP CRBE



- resolution: 10ms



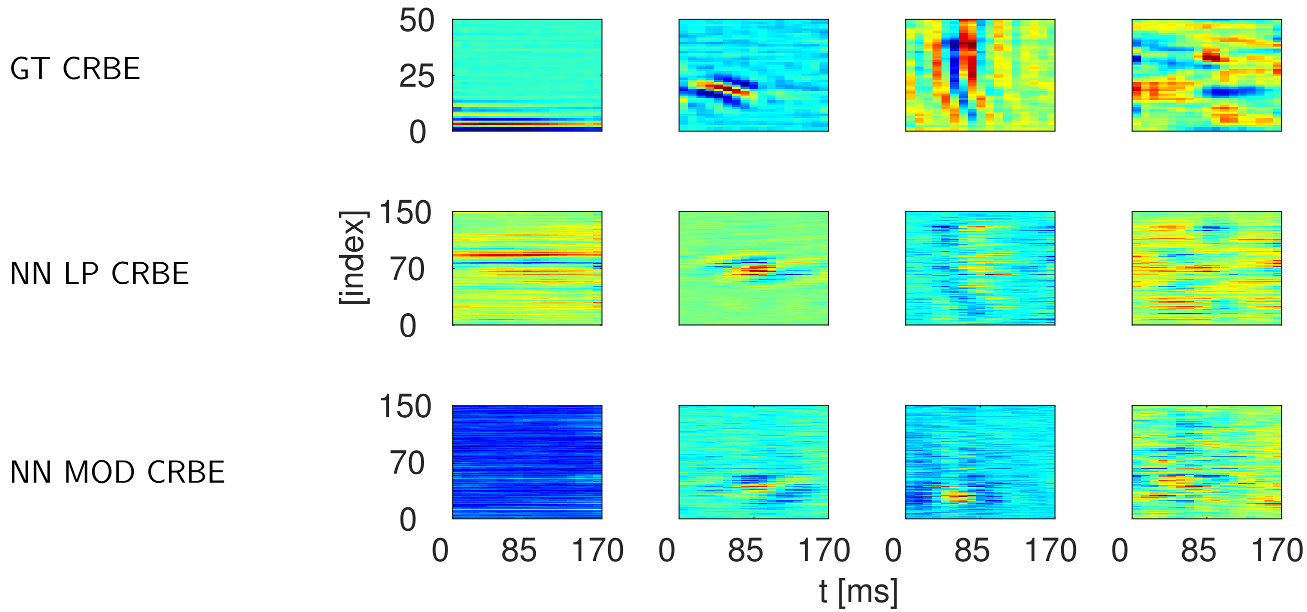
NN MOD CRBE

- resolution: 10ms

Weight analysis

Analyzing the first layer of the back-end:

- X_t contains 17 frames of multi-resolution spectra.
- Selecting weights belonging to a specific spectral representation.
- Plotting in 2D: filter frequency and position in the time-window.
  - GT front-end: 50x17 patches.
  - NN front-end: 150x17, using estimated center frequencies of TF filter.



- Frequency selectors, Gabor patches, delta features, complex CRBE patterns.

Conclusions

- Direct waveform model could match the performance of optimized cepstral features, using less than 250 hours of speech.
- Still, slight gap between hand-crafted and data-driven features after segment-level normalization.
- The data-driven front-end strongly depends on the back-end, less portable.

- NN based signal processing prefers to learn modulation spectral representation.
  - For higher resolution in modulation frequency, the envelop filter response should be up to 1 sec long.

- Weight analysis reveals patterns similar to activations in the auditory cortex.

Thank you for your attention

Any questions?

References

Davis, S. and Mermelstein, P. (1980).
Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.
IEEE Trans. on Acoustics, Speech, and Signal Processing, 28(4):357–366.

Fletcher, H. and Munson, W. A. (1933).
Loudness, its definition, measurement and calculation.
The Journal of the Acoustical Society of America, 82(5):82–108.

Furui, S. (1981).
Comparison of speaker recognition methods using statistical features and dynamic features.
IEEE Trans. on Acoustic, Speech, and Signal Processing, 29(3):342–350.

Ghahremani, P., Manohar, V., Povey, D., and Khudanpur, S. (2016).
Acoustic modelling from the signal domain using CNNs.
In *Interspeech*, pages 3434–3438.

Golik, P., Tüske, Z., Schlüter, R., and Ney, H. (2015).
Convolutional neural networks for acoustic modeling of raw time signal in LVCSR.
In *Interspeech*, pages 26–30.

Hermansky, H. (1990).
Perceptual linear predictive (PLP) analysis of speech.
Journal of the Acoustical Society of America, 87(4):1738–1752.

Hoshen, Y., Weiss, R. J., and Wilson, K. W. (2015).
Speech acoustic modeling from raw multichannel waveforms.
In *ICASSP*, pages 4624–4628.

Conclusions

Li, B., Sainath, T. N., Weiss, R. J., Wilson, K. W., and Bacchiani, M. (2016).
Neural network adaptive beamforming for robust multichannel speech recognition.
In *Interspeech*.

Palaz, D., Collobert, R., and Magimai.-Doss, M. (2013).
Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks.
In *Interspeech*, pages 1766–1770.

Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., and Vinyals, O. (2015).
Learning the speech front-end with raw waveform CLDNNs.
In *Interspeech*, pages 1–5.

Schlüter, R., Bezrukov, I., Wagner, H., and Ney, H. (2007).
Gammatone features and feature combination for large vocabulary speech recognition.
In *ICASSP*, pages 649–652.

Tüske, Z., Golik, P., Schlüter, R., and Ney, H. (2014).
Acoustic modeling with deep neural networks using raw time signal for LVCSR.
In *Interspeech*, pages 890–894.

von Békésy, G. (1960).
Experiments in Hearing.
McGraw-Hill, New York.

Zhu, Z., Engel, J. H., and Hannun, A. (2016).
Learning multiscale features directly from waveforms.
In *Interspeech*, pages 1305–1309.