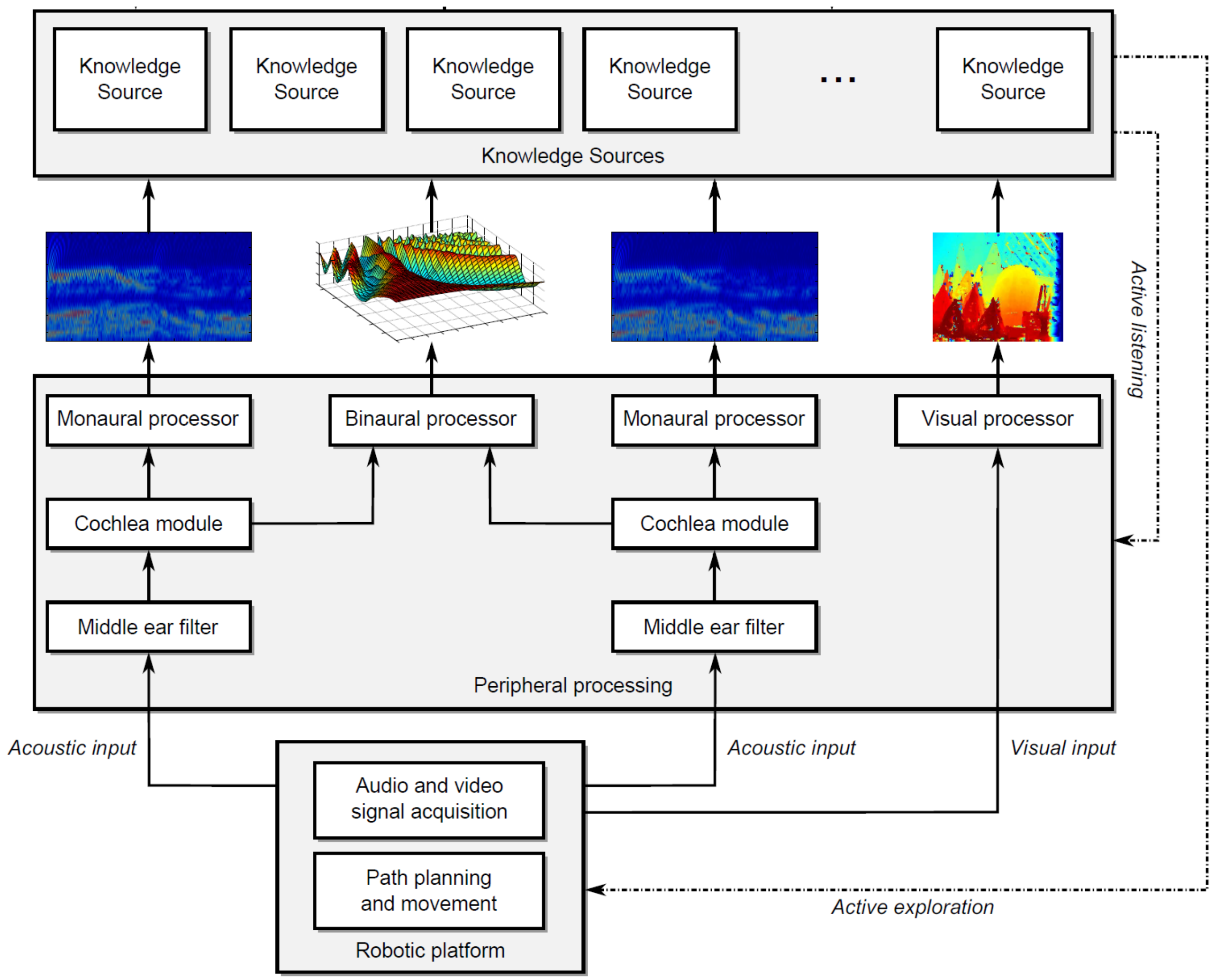
Sound Type Classification using Deep Neural Networks

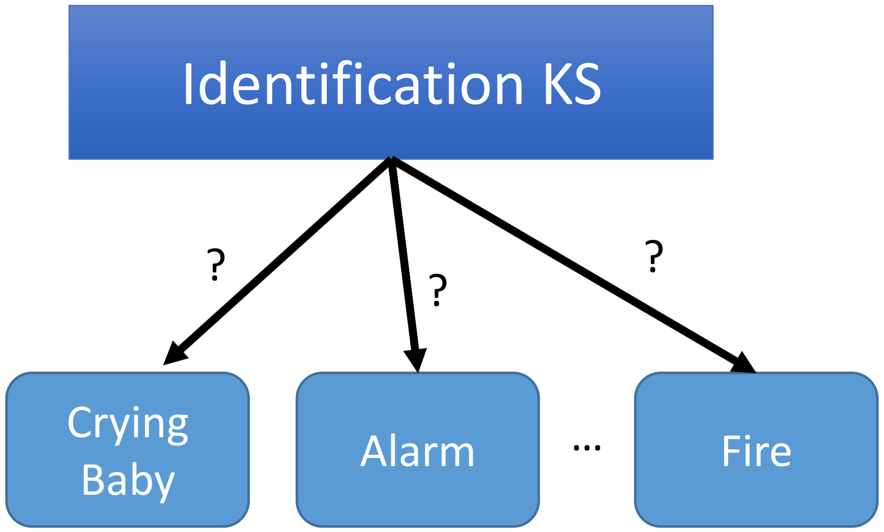
Johannes Mohr

Group Talk

April 22, 2016

The Two!Ears-System





NIGENS (NI General Sound) database

- 11 classes of everyday sounds: engine, crash, footsteps, piano, dog, phone, knock, fire, crying baby, alarm, female speech
  - 50 WAV files per class
- 1 general sound class
  - 237 WAV files (not including sounds from the other classes)










Sound Type Classification

- Current identification KS in Two!Ears:
  - Rate maps, spectral features, AMS features, onset strengths
  - Features based on statistics of time series and derivatives
  - Classification using Lasso or SVMs
- Deep Neural Networks
  - Automatically extract “useful features”
  - Might be better at handling overlaid sounds
  - Inputs: raw rate maps and amplitude modulation spectra
  - Compare to current identification KS using only features derived from the raw rate maps /AMS spectra

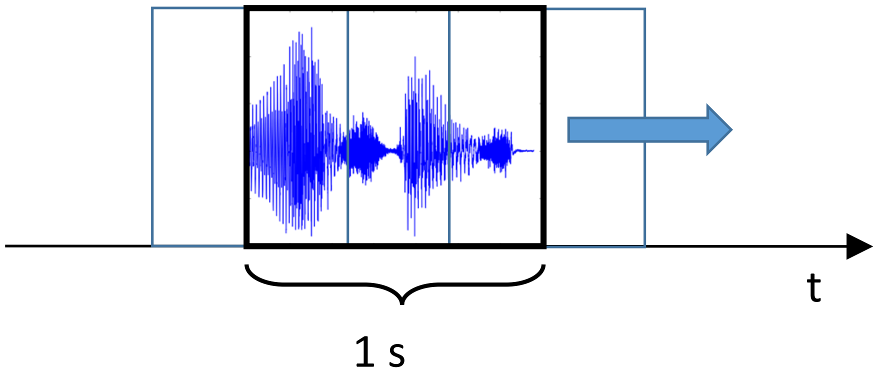
Identification Pipeline

Sound aqcuisition

Audio Signal Block of 1 s

Sound data from room simulator or robot

- Sound Amplitude as function of time
- Sampled at 44.1 kHz
- Time window of 1 s at “jumps” of 1/6 s



Identification Pipeline

Sound aqcuisition → Auditory Frontend

Audio Signal Block of 1 s

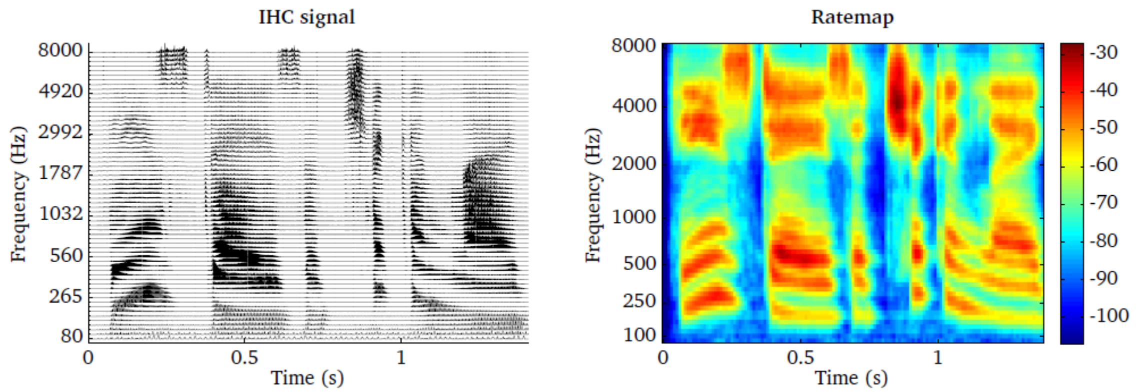
Rate Map

Amplitude Modulation Spectrogram

Sound data from room simulator or robot

Representations

- Rate Map
  - auditory spectrograms that represent auditory nerve firing rates for each of 63 timeframes (20 ms) and each of 16 individual gammatone frequency channels
  - computed by smoothing the corresponding inner hair cell signal representation with a leaky integrator



Identification Pipeline

Sound aqcuisition → Auditory Frontend

Audio Signal Block of 1 s

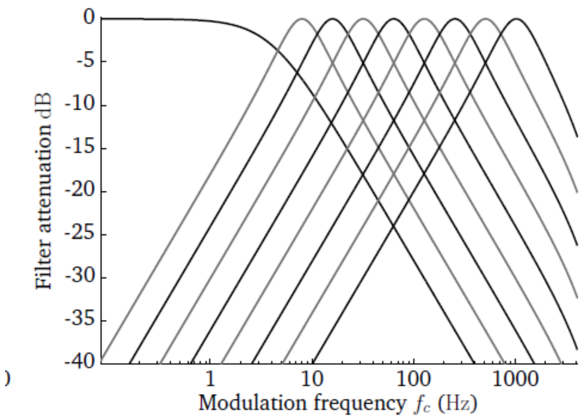
Rate Map

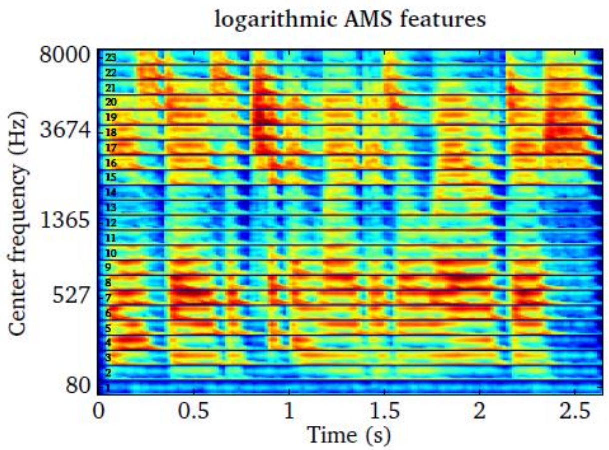
Amplitude Modulation Spectrogram

Sound data from room simulator or robot

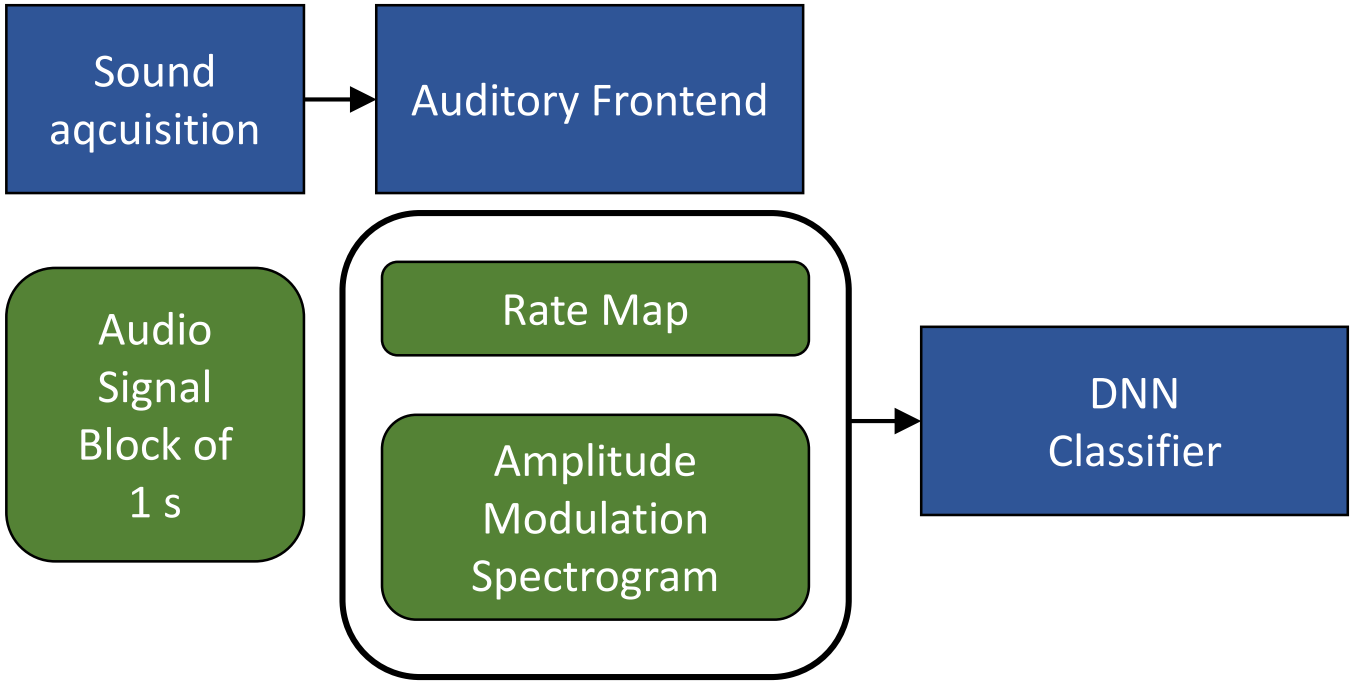
Representations

- Amplitude Modulation Spectrogram
  - each frequency channel of the inner hair cell representation is analysed by a bank of logarithmically-scaled modulation filters
  - for each of 63 time frames there are 16 x 9 values (16 frequency channels, 9 modulation filters)





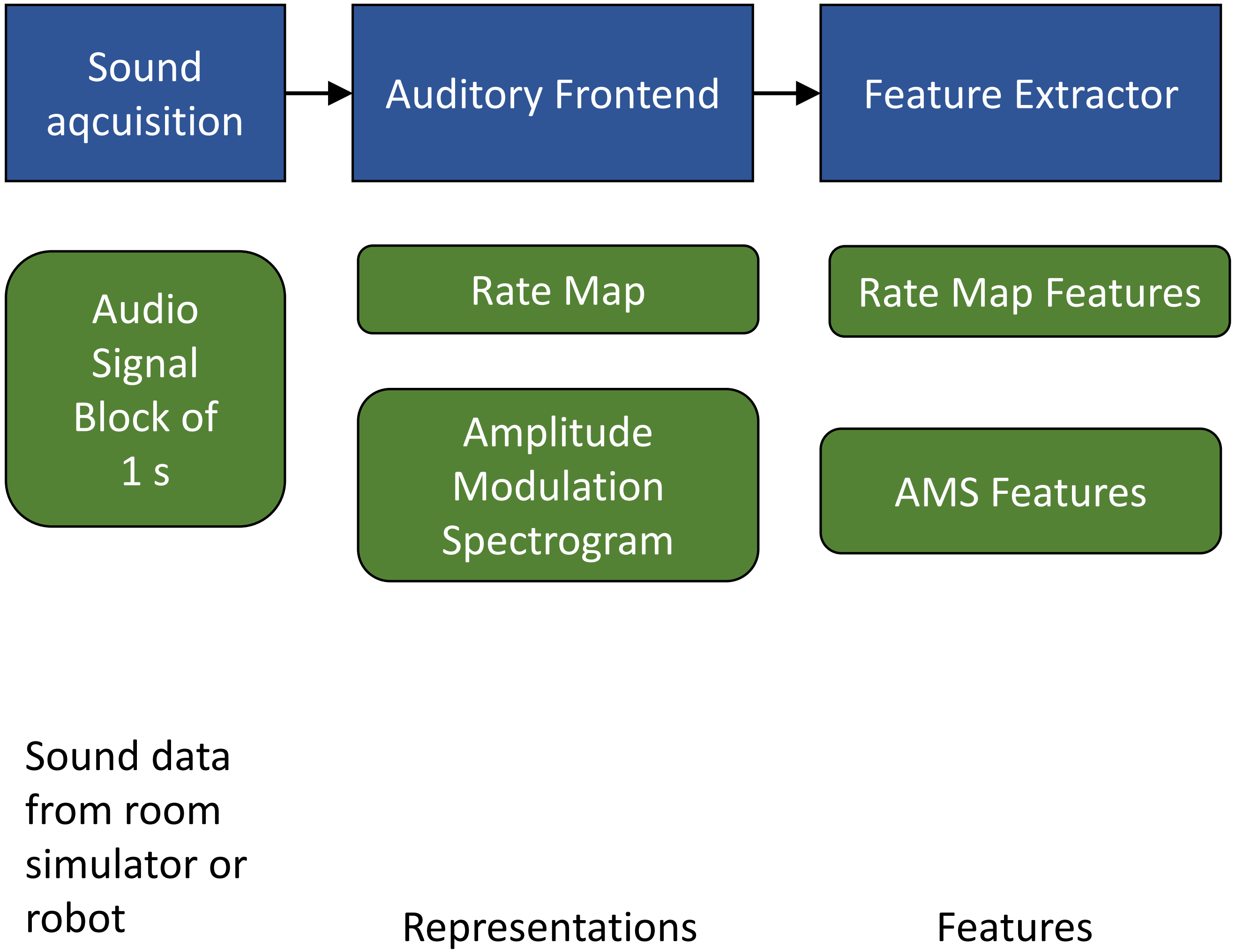
Input to the Deep Neural Network



Sound data from room simulator or robot

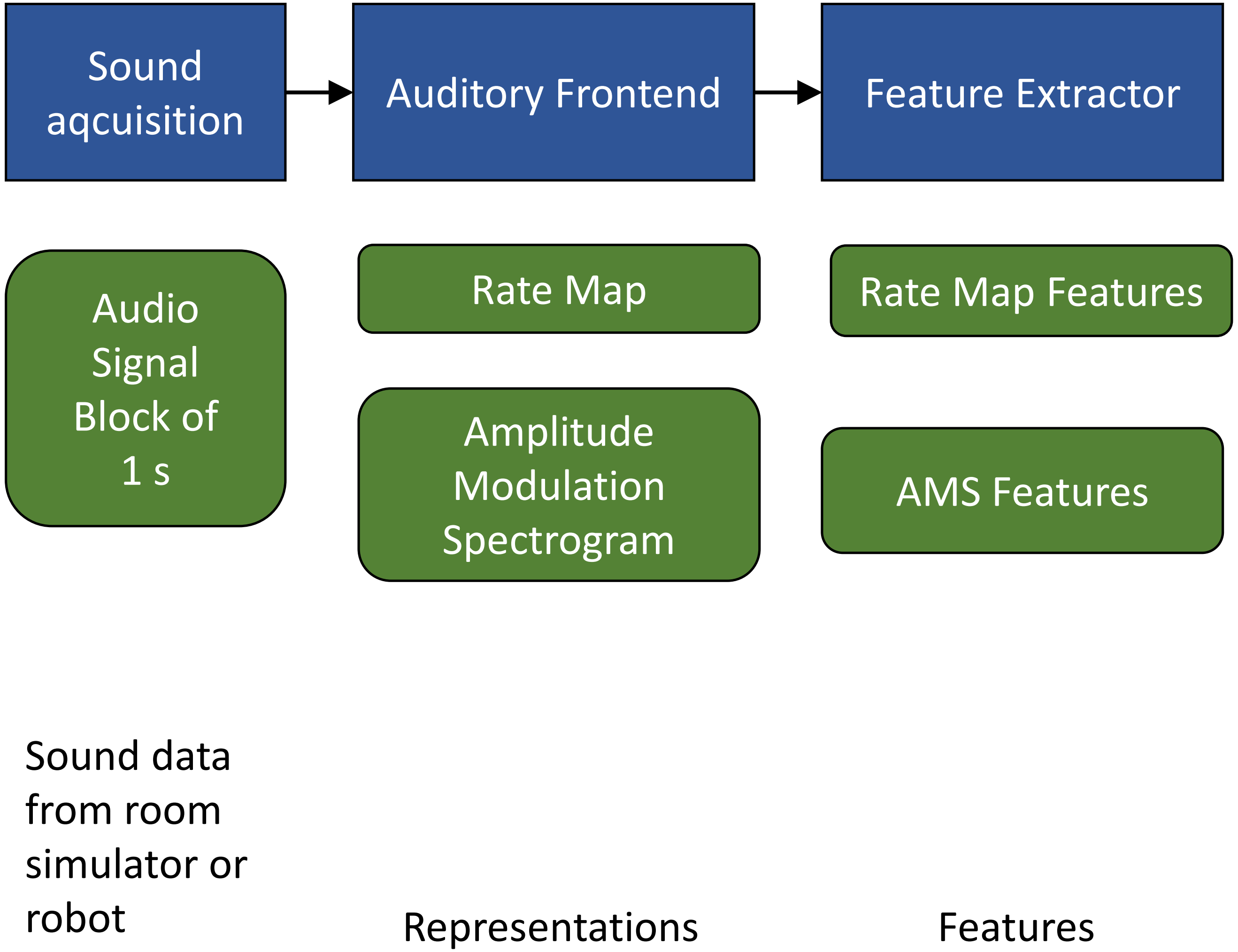
Representations

Identification Pipeline



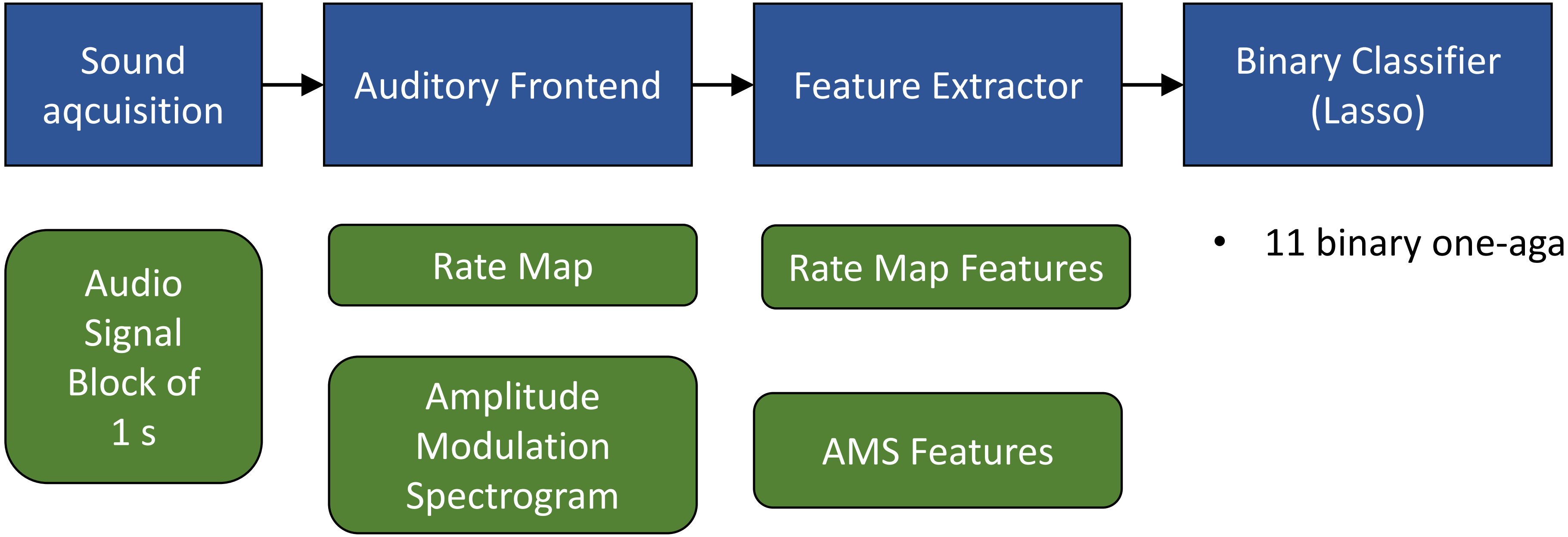
- For each representation, features are extracted by averaging over both channels and calculating
  - Sample L-moments (L-mean, L-scale, L-skewness, L-kurtosis) of the representation over the frames in the block
  - Ratemaps: 1., 2. and 3. L-moment
  - AMS: 1. and 2. L-moment

Identification Pipeline



- Feature Set rm+ams (336 features)
  - 48 (3x16) rate map features
  - 288 (2x9x16) amplitude modulation spectrogram features
- Feature Set rm (48 features)
  - 48 (3x16) rate map features

Identification Pipeline



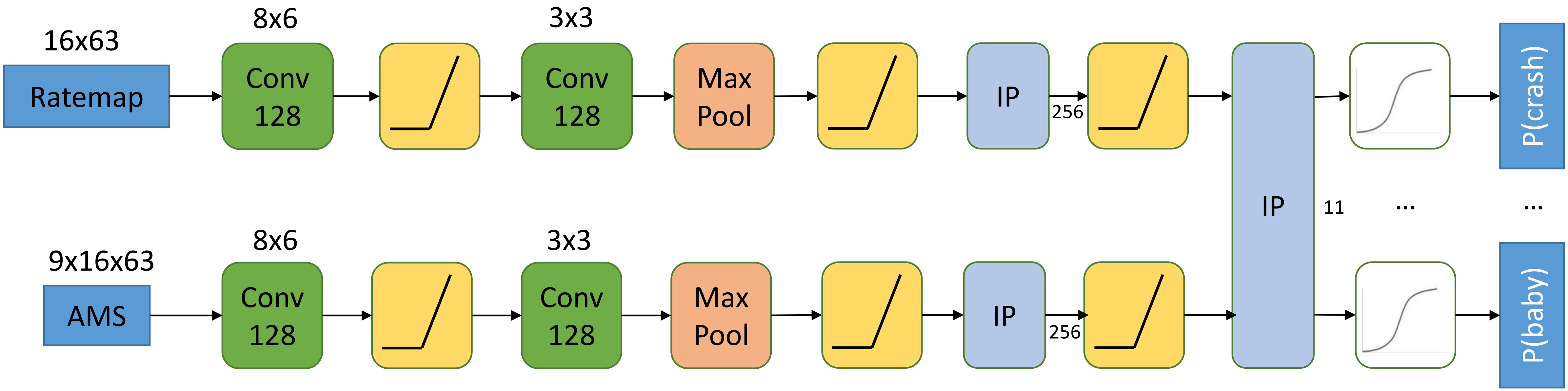
- 11 binary one-against-all classifiers

Sound data from room simulator or robot

Representations

Features

Deep Neural Network Model 1
(sigmoid output, cross-entropy loss)



Cross-Entropy Loss

- Cross entropy:

$$H(p,q) = -\sum_{y} p(y)\log q(y)$$

- Cross-entropy loss:

$$E = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c} p_{cn}\log o_{cn}\,(1-p_{cn}) + \log(1-o_{cn}\,)$$

  - o_{cn} is the output of the network for class c and example n
  - p_{cn} is 1 if the true label of example n is c, and 0 otherwise
  - N is the sample size

Regularization of deep neural networks

- early stopping/ annealed learning rate
- L1 weight penalty
- L2 weight penalty (weight decay)

$$\Delta w_i(t+1) = w_i - \eta \frac{\partial E}{\partial w_i} - \eta \lambda w_i$$

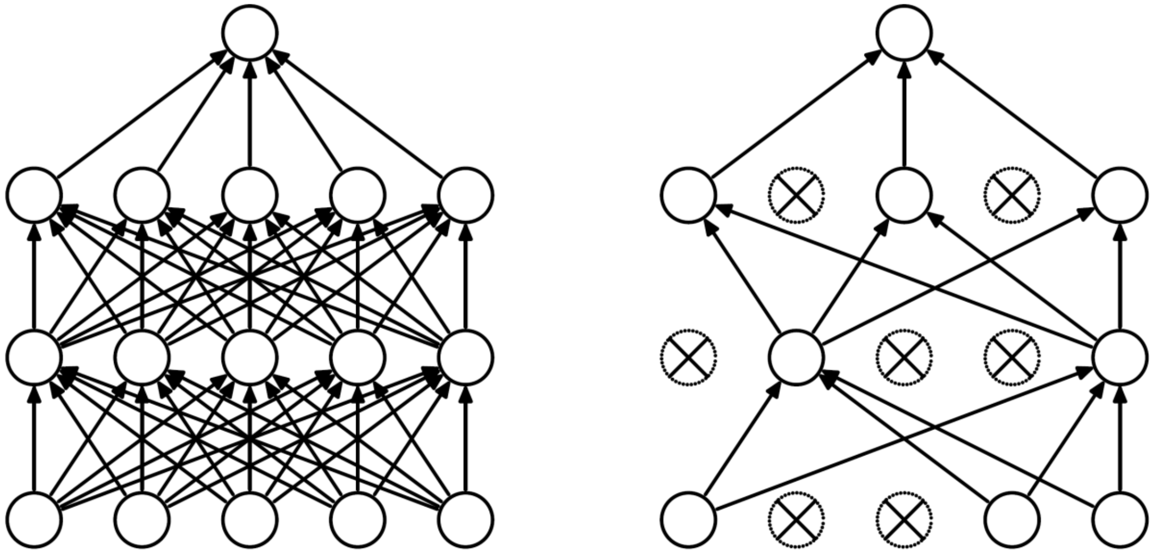
- momentum term

$$\Delta w_i(t+1) = w_i - \eta \frac{\partial E}{\partial w_i} + \alpha \Delta w_i(t)$$

- soft weight sharing (co-adapt GMM for weight vectors)
- posterior weighted average of predictions over all possible parameter settings

Dropout

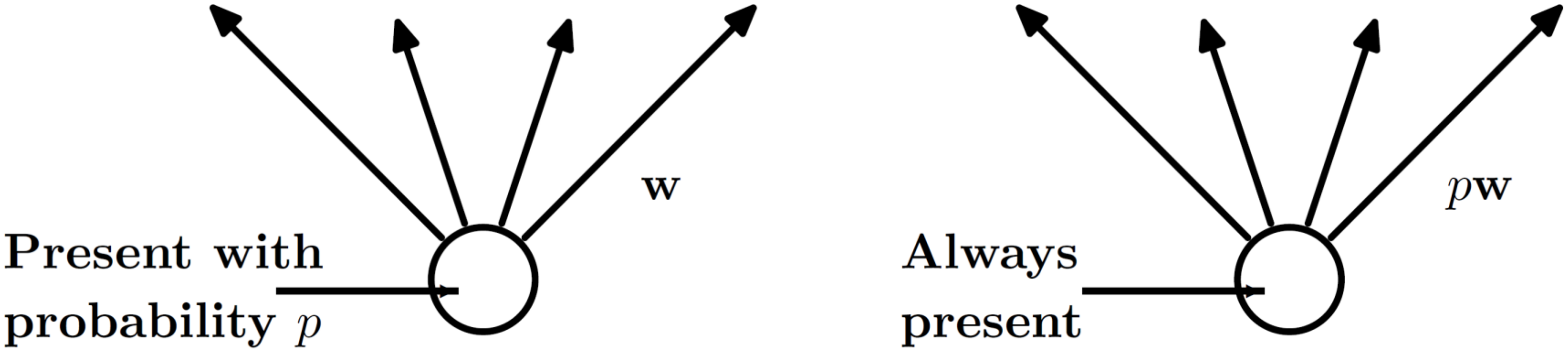
- Drop units from the neural network during training



- Training neural network with n units and dropout $\cong$ Training a collection of 2^n possible thinned networks with shared weights

Dropout

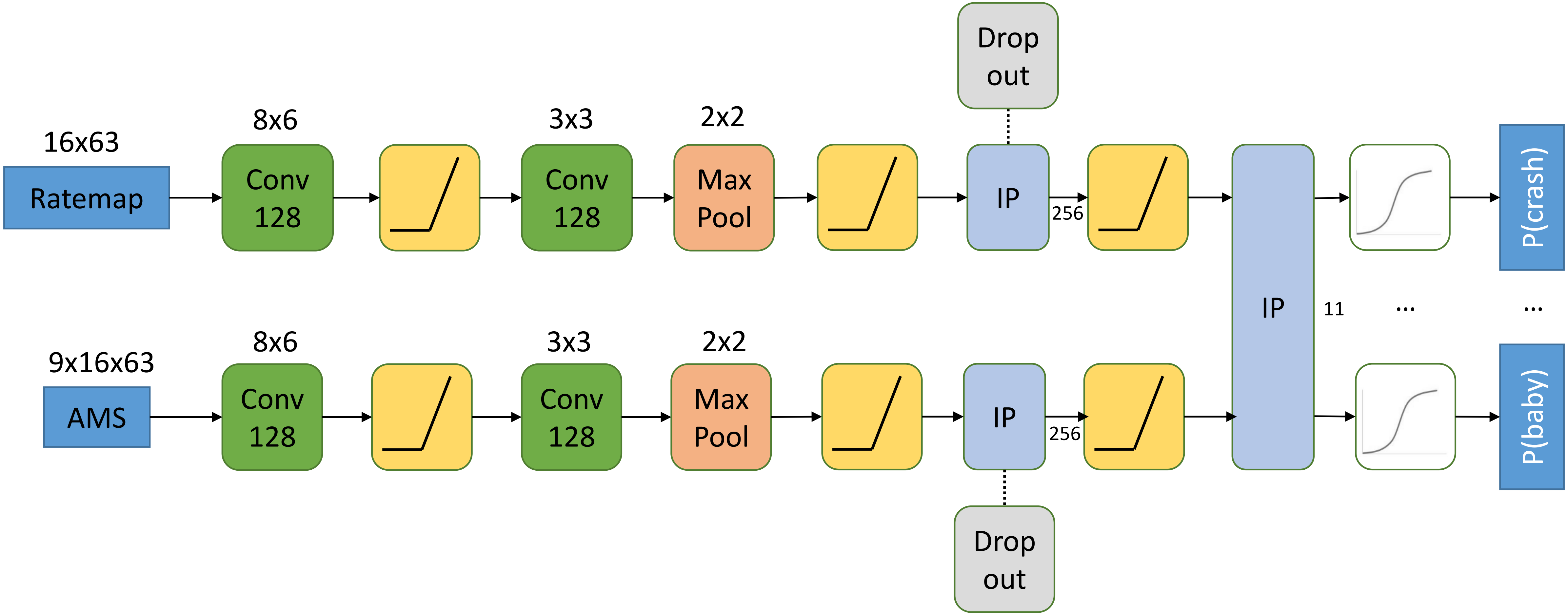
- No dropout at test time, but scale down weights



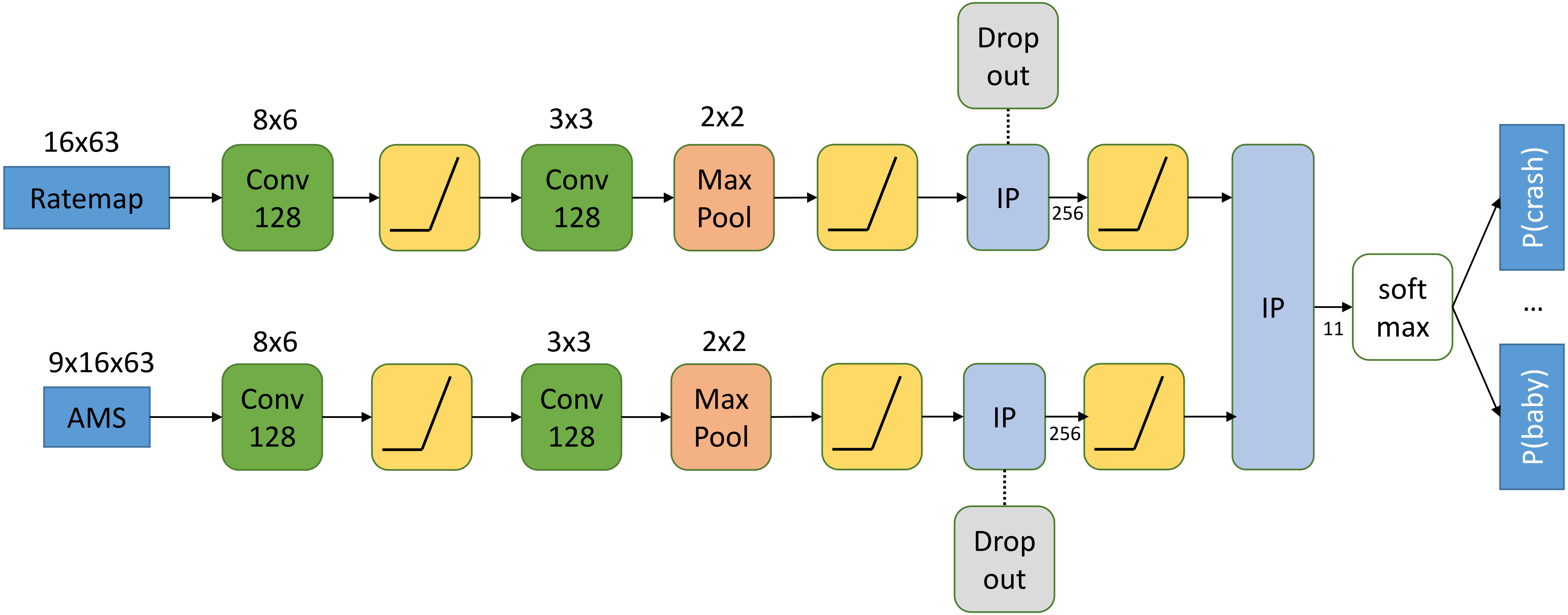
(a) At training time (b) At test time

- Expected output at training time is the same as the output at test time

Deep Neural Network Model 1
(sigmoid output, cross-entropy loss)



Deep Neural Network Model 2
(soft max output, multinomial logistic loss)

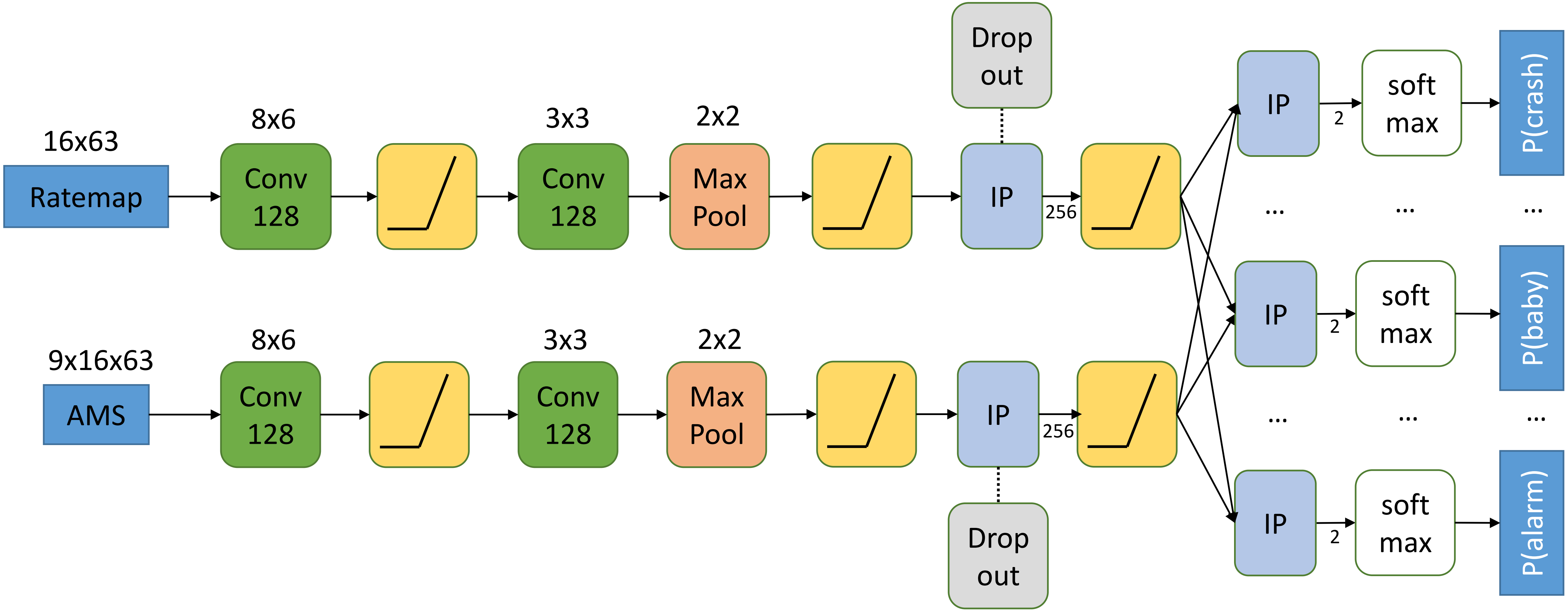


Multinomial Logistic Loss

$$E = -\frac{1}{N}\sum_{n} \log o_{kn}$$

- o_{kn} is the output of the softmax layer of the network for the true class k of example n, where $\sum_k \log o_{kn} = 1$
- N is the sample size

Deep Neural Network Model 3 (one-against-all)



Experiments

- Auditory Scenes generated by Binaural Simulator
    - Anechoic sounds, no overlay
- Sound files Split into a training (75%) and a test set (25%)
- Lasso: binary classifiers trained for all 11 sound types
- DDNs
    - Training set balanced over all classes
    - Stochastic gradient descent
        - weight decay (λ = 0.0005), high momentum (α = 0.99), base learning rate η = 0.0001, stepwise reduction of learning rate by γ = 0.1 every 30000 iterations
- performance evaluation: balanced accuracy of each sound type

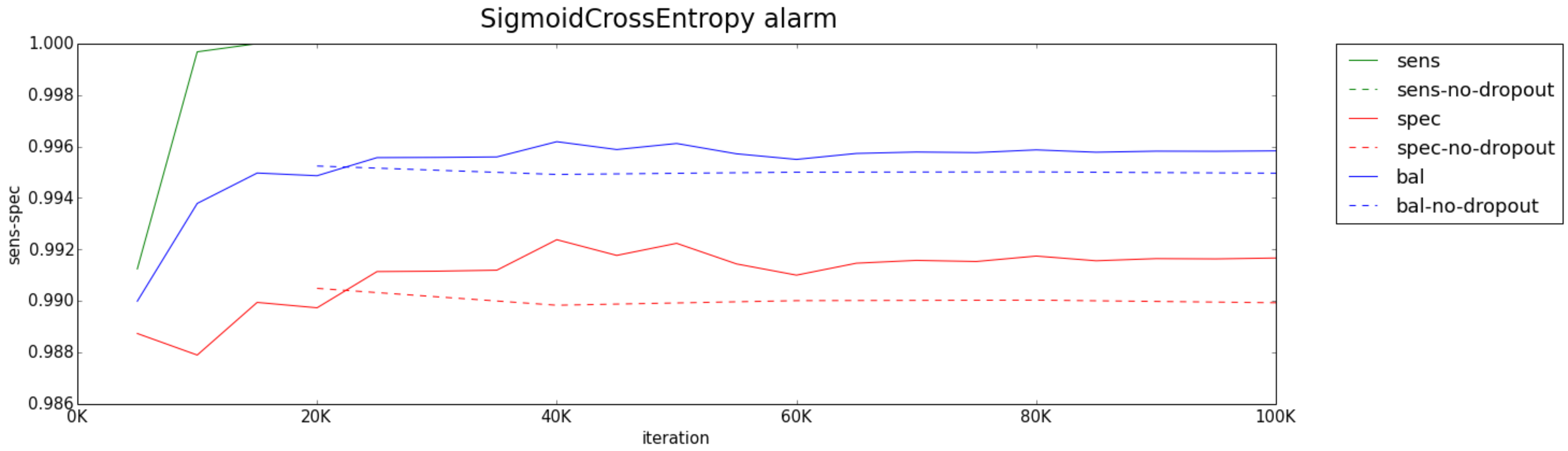
Results

	Lasso (RM + AMS)					DNN (Sigmoid Cross Entropy)			DNN (Softmax)			DNN (1-vs-all)		
	Sens	Spec	BalAcc	perf	(Full FS)	Sens	Spec	BalAcc	Sens	Spec	BalAcc	Sens	Spec	BalAcc
fem. speech	1.00	1.00	1.00	1.00	1.00	0.97	0.99	0.98	0.96	1.00	0.98	0.95	1.00	0.98
alarm	0.95	0.87	0.91	0.90	0.92	**0.31**	0.96	**0.64**	**0.32**	1.00	**0.66**	**0.18**	1.00	**0.59**
baby	0.96	0.96	0.96	0.96	0.98	0.99	0.95	0.97	0.96	0.98	0.97	0.95	1.00	0.97
fire	0.95	0.95	0.95	0.95	0.94	0.98	0.93	0.95	0.96	0.98	0.97	0.94	0.99	0.97
knock	0.99	0.99	0.99	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
phone	0.99	0.94	0.97	0.96	0.94	0.93	0.99	0.96	**0.73**	1.00	**0.86**	**0.63**	1.00	**0.82**
dog	0.98	0.96	0.97	0.97	0.99	0.95	0.97	0.96	0.94	1.00	0.97	**0.87**	1.00	**0.94**
piano	0.83	0.93	0.88	0.87	0.97	0.97	0.96	0.96	0.81	0.99	0.90	**0.68**	0.99	**0.84**
footsteps	1.00	0.96	0.98	0.97	0.94	0.99	0.90	0.94	0.98	0.99	0.98	0.93	0.99	0.96
crash	0.91	0.84	0.88	0.87	0.88	0.83	0.91	0.87	0.76	0.96	0.86	0.83	0.91	0.87
engine	0.66	0.82	0.74	0.73	0.83	0.79	0.86	0.82	**0.42**	0.94	**0.68**	**0.42**	0.94	**0.68**

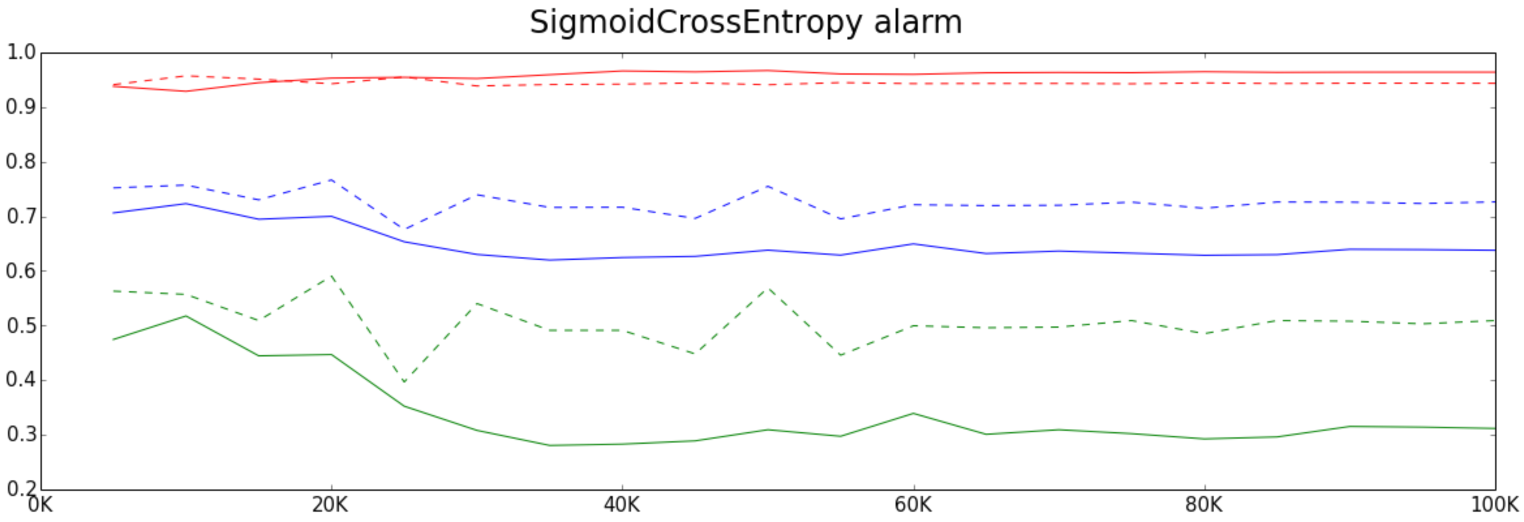
$$\text{perf} = 1 - \sqrt{((1 - \text{sensitivity})^2 + (1 - \text{specificity})^2)/2)}$$

Alarm

Training



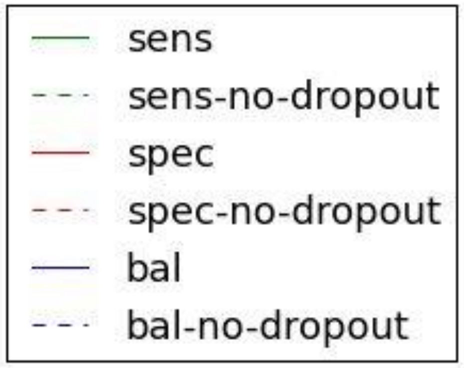
Test

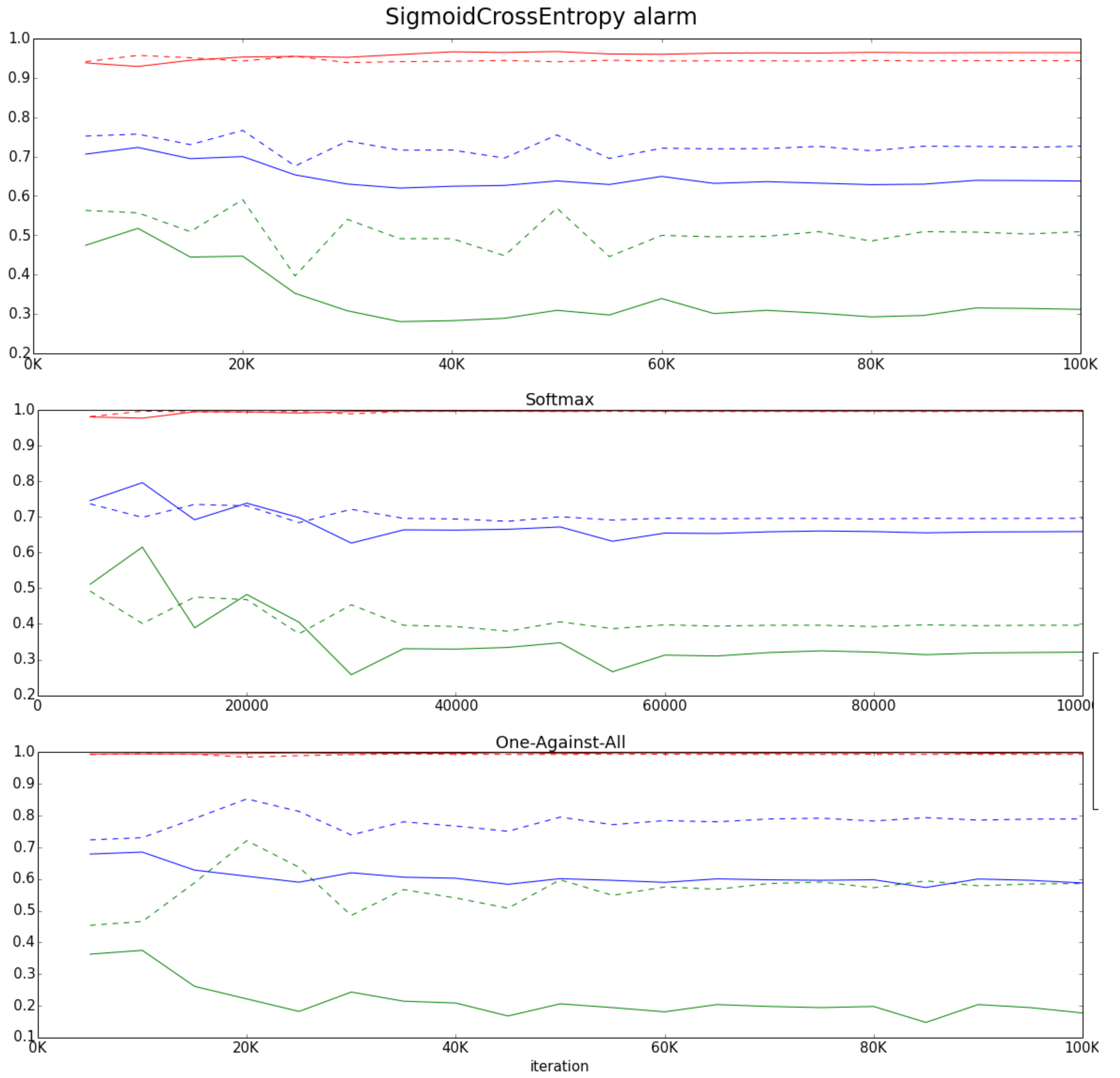


Alarm

Performance on Test Set

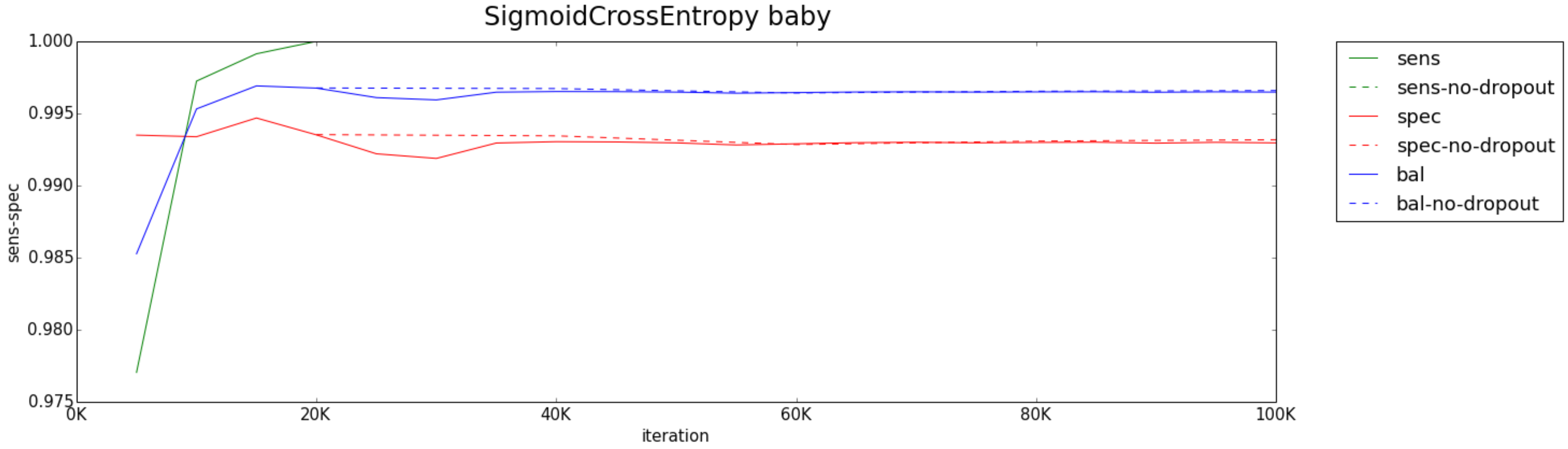
	Sens.	Spec.	BalAcc
Lasso (RM+AMS)	**0.95**	**0.87**	**0.91**
Sigmoid cross-entropy	0.31	0.96	0.64
Sigmoid cross-entropy (nd)	0.51	0.94	0.73
Softmax	0.32	1.00	0.66
Softmax (nd)	0.40	1.00	0.70
One-against-all	0.17	1.00	0.59
One-against-all (nd)	0.59	1.00	0.79



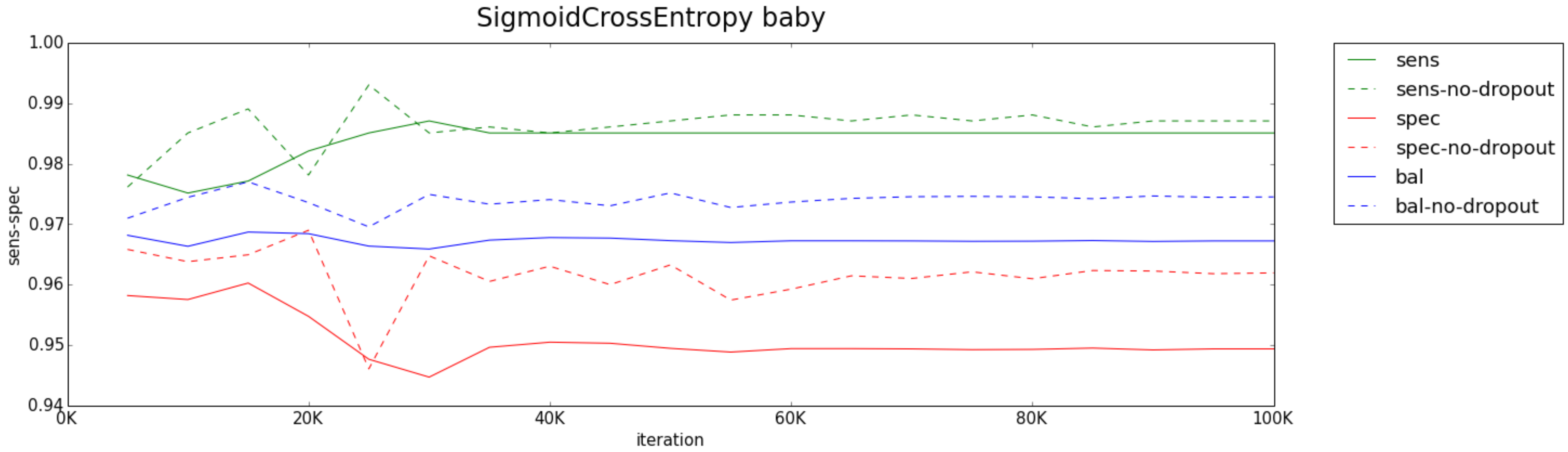


Baby

Training



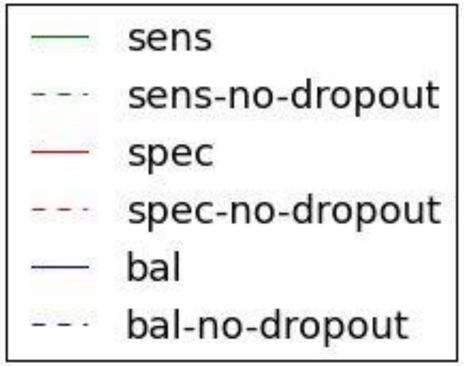
Test

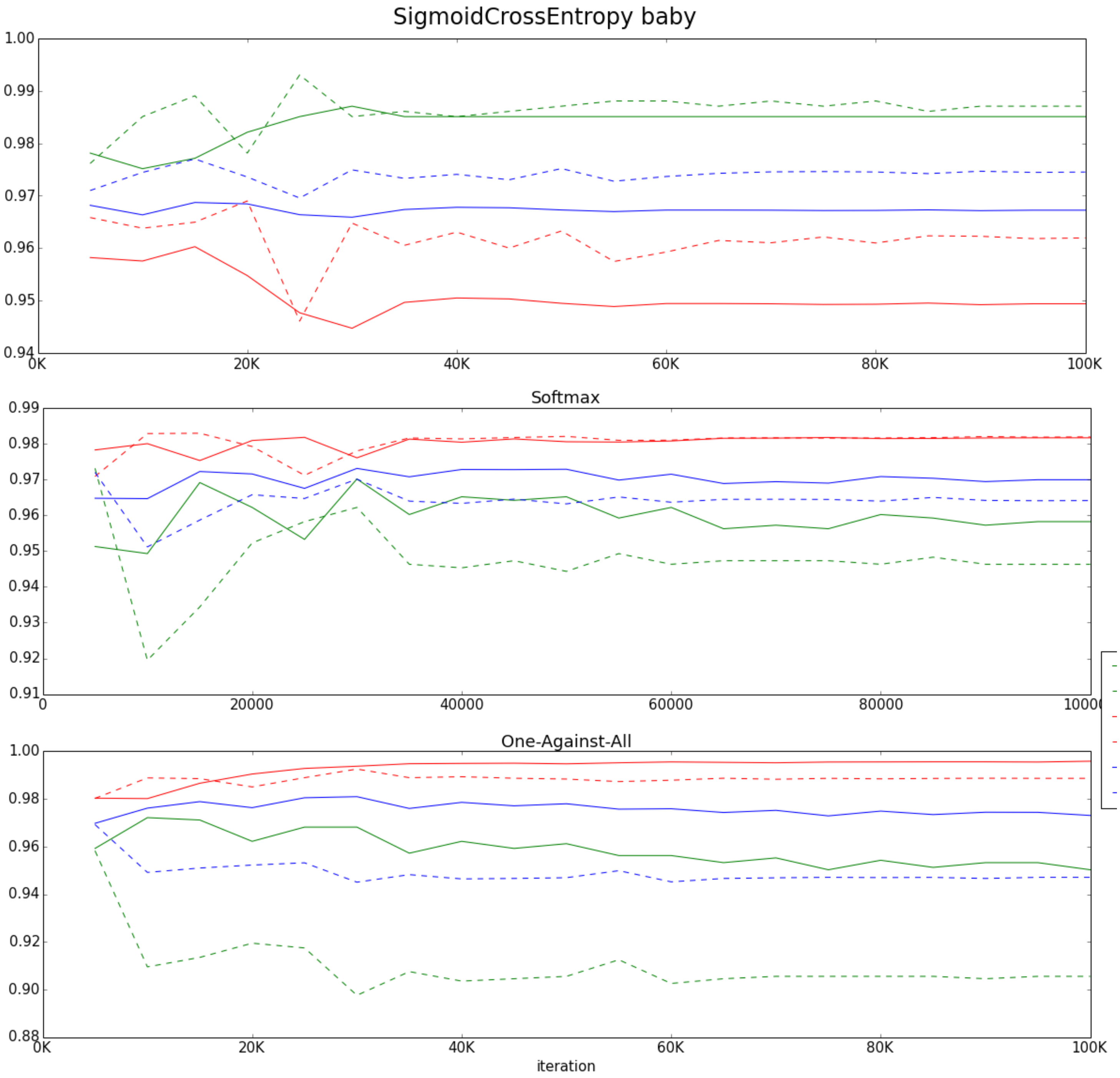


Baby

Performance on Test Set

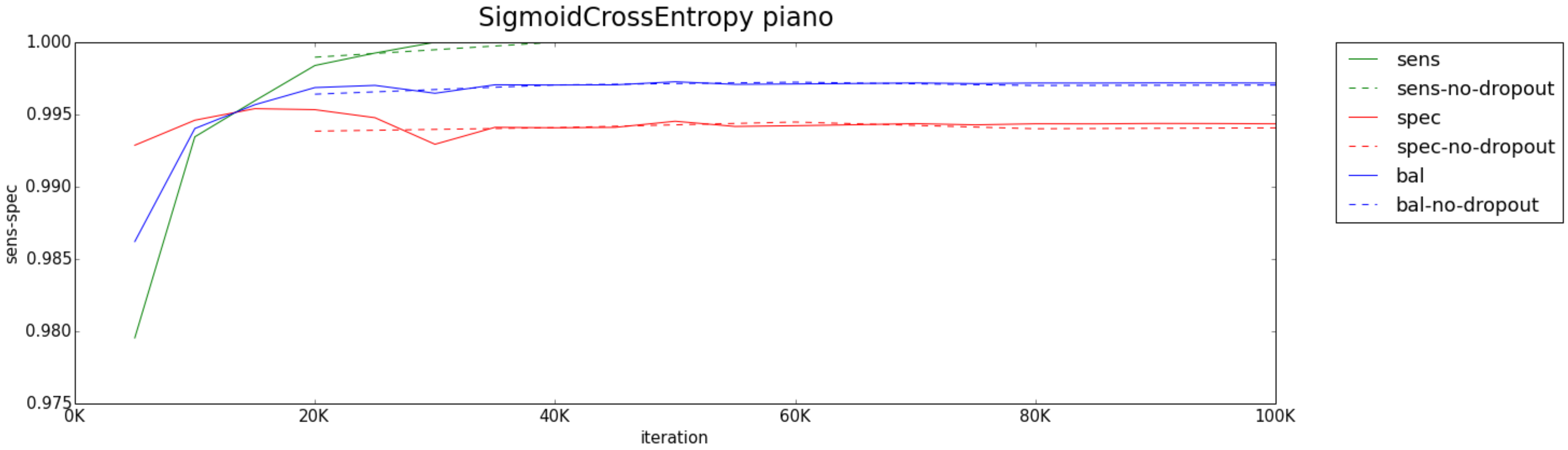
	Sens.	Spec.	BalAcc
Lasso (RM+AMS)	0.96	0.96	0.96
Sigmoid cross-entropy	0.99	0.95	0.97
Sigmoid cross-entropy (nd)	0.99	0.96	0.97
Softmax	0.96	0.98	0.97
Softmax (nd)	0.95	0.98	0.96
One-against-all	0.95	1.00	0.97
One-against-all (nd)	0.91	0.99	0.95



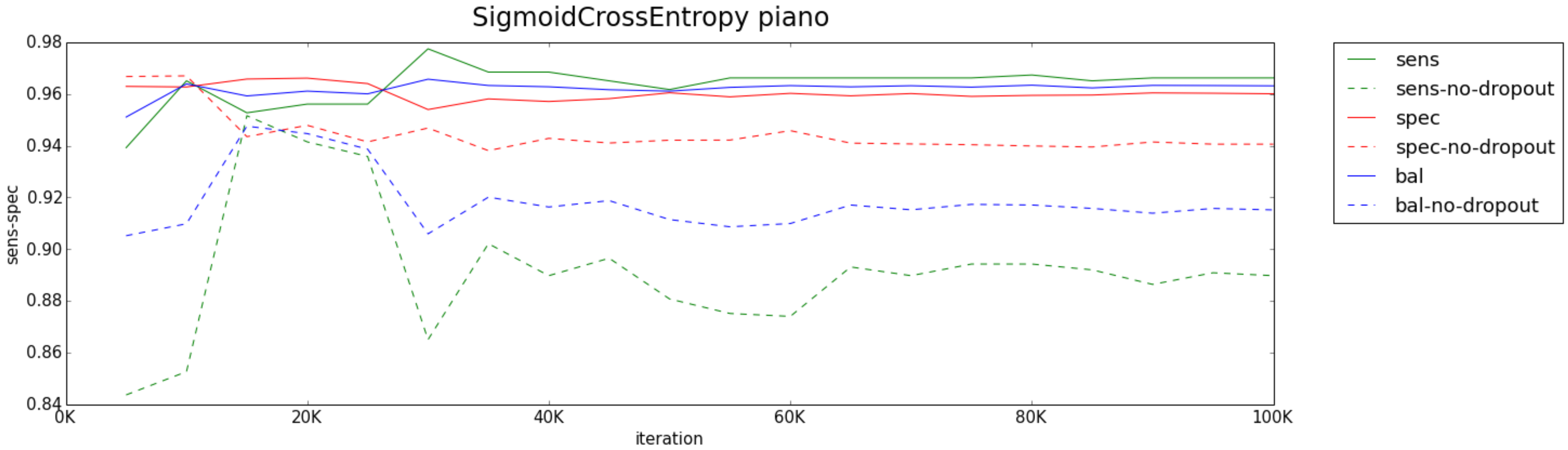


Piano

Training



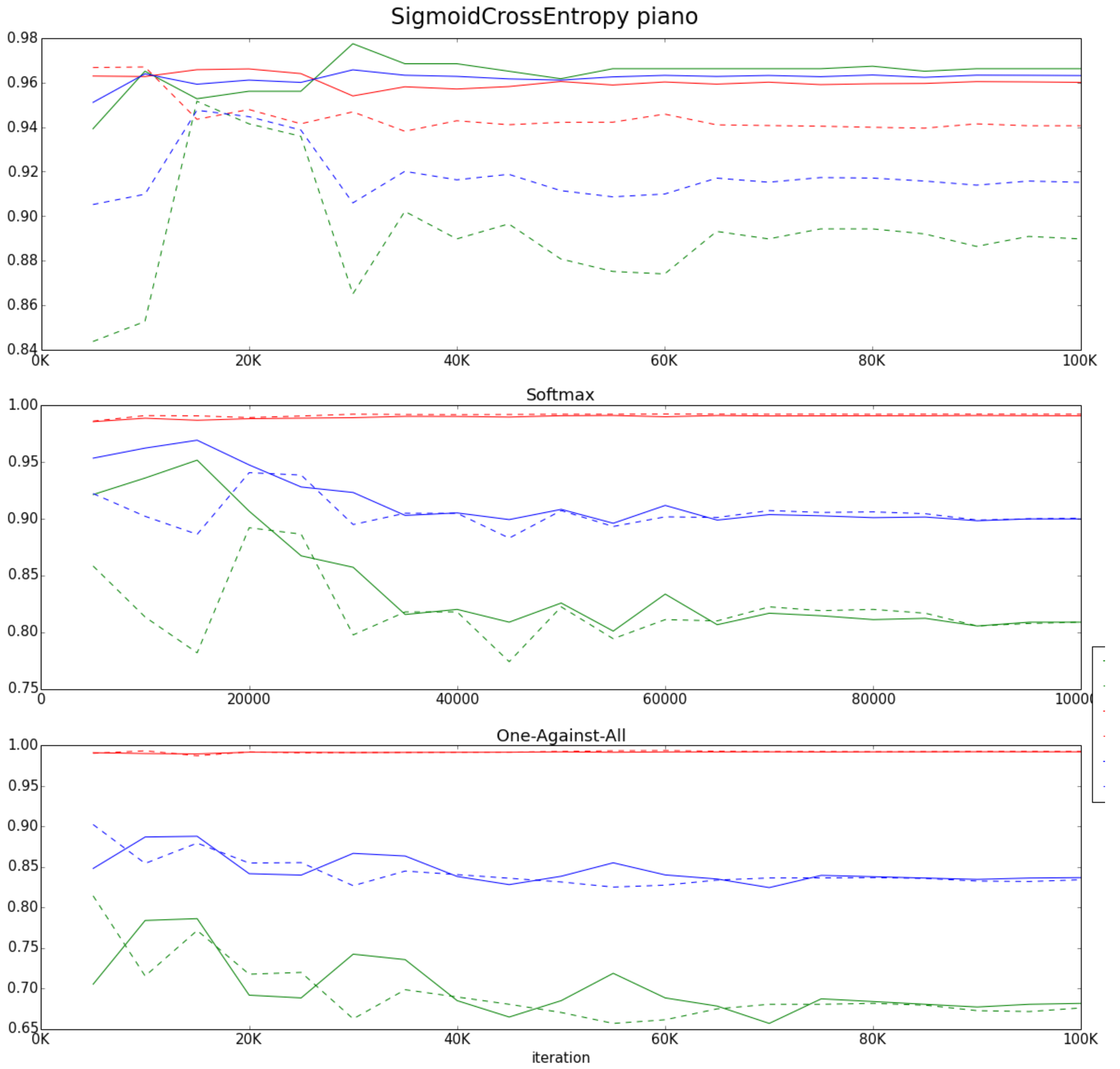
Test

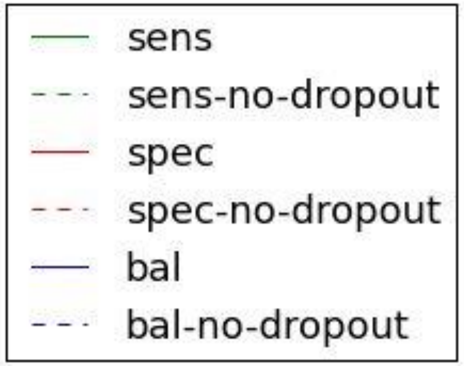


Piano

Performance on Test Set

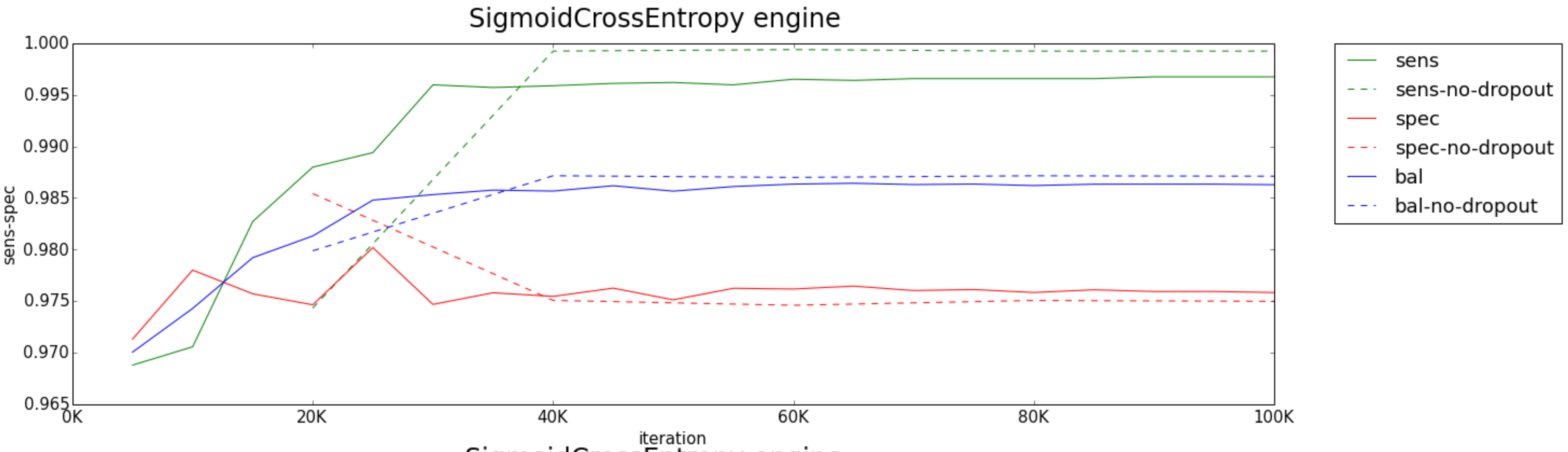
	Sens.	Spec.	BalAcc
Lasso (RM+AMS)	0.83	0.93	0.88
Sigmoid cross-entropy	**0.97**	**0.96**	**0.96**
Sigmoid cross-entropy (nd)	0.89	0.94	0.92
Softmax	0.81	0.99	0.90
Softmax (nd)	0.81	0.99	0.90
One-against-all	0.68	0.99	0.83
One-against-all (nd)	0.68	0.99	0.83



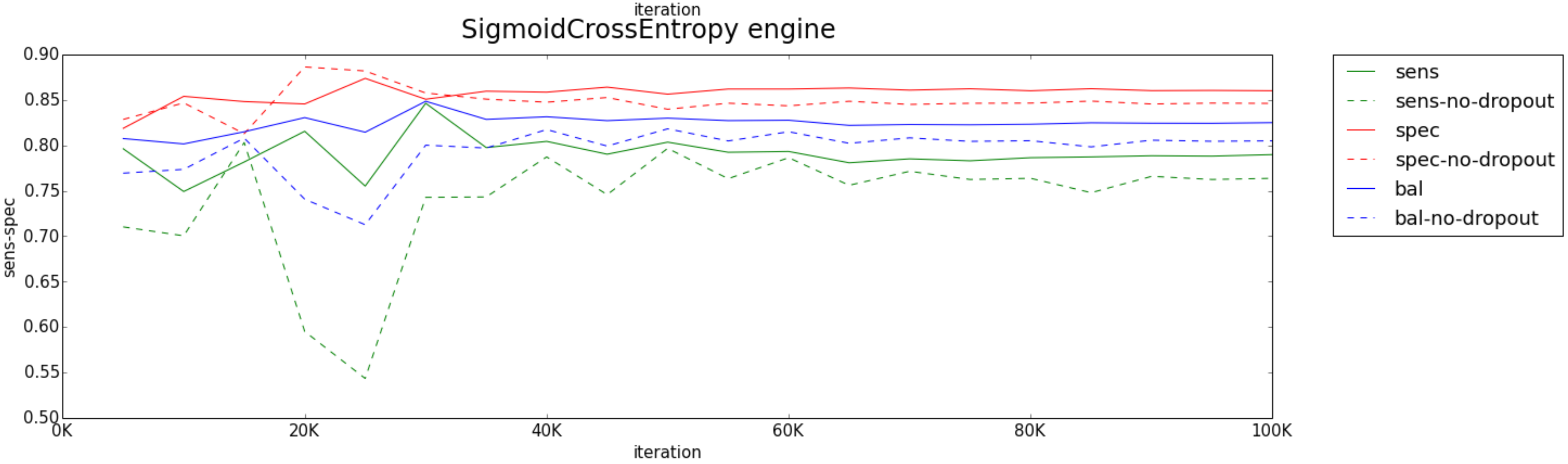


Engine

Training



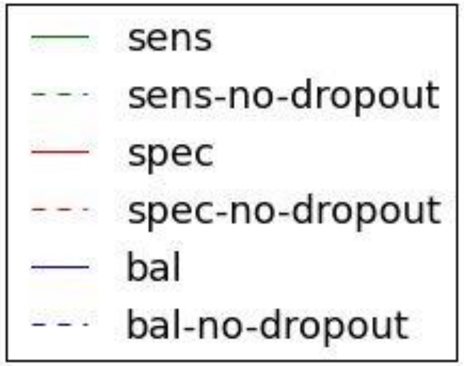
Test

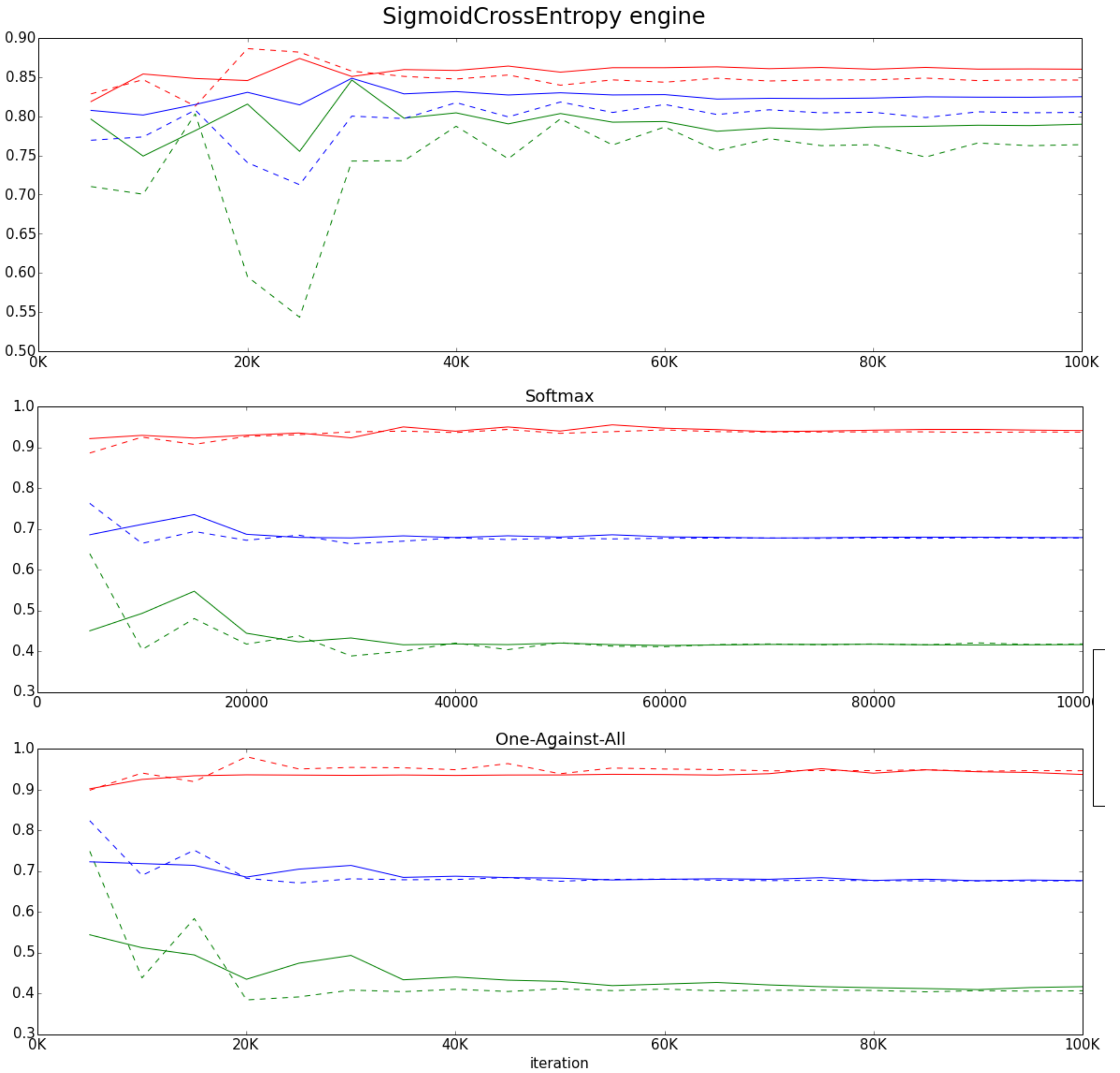


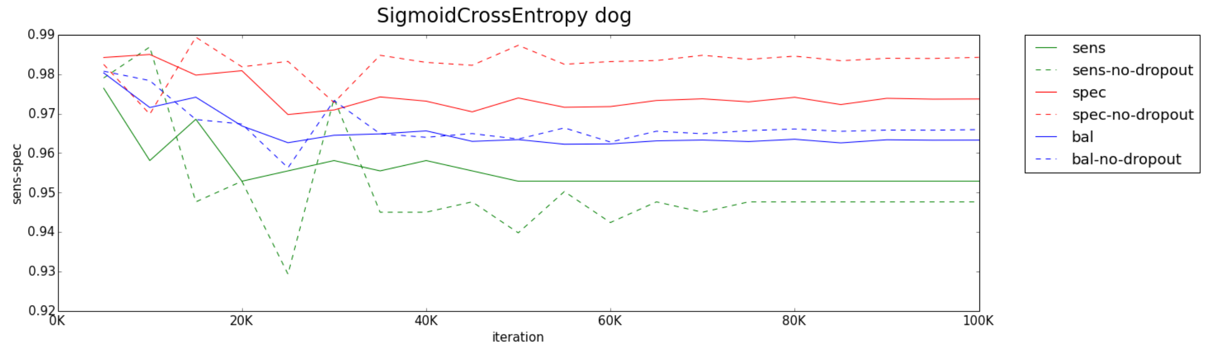
Engine

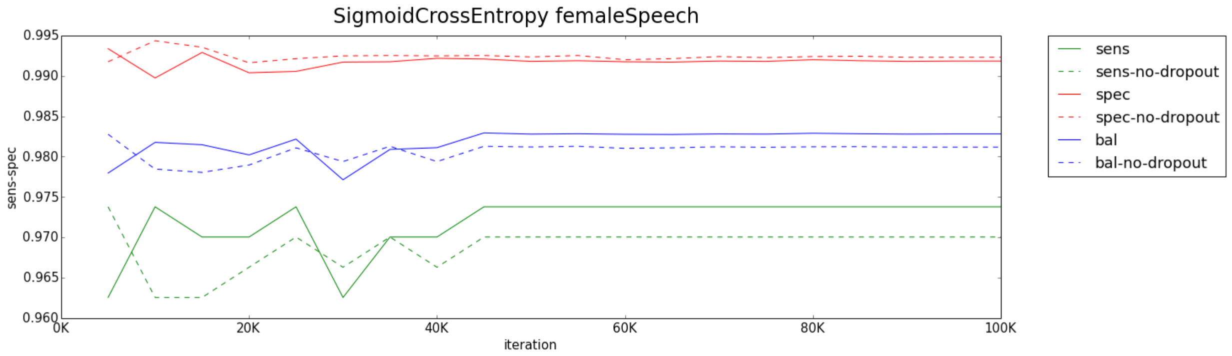
Performance on Test Set

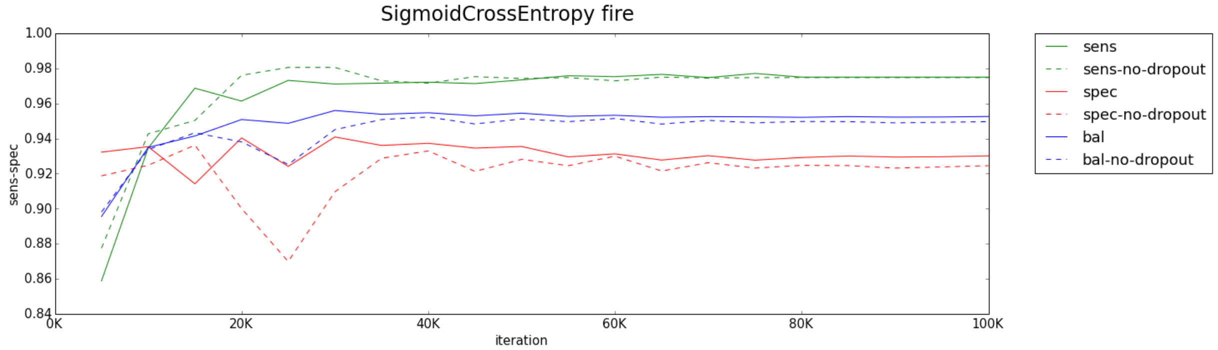
	Sens.	Spec.	BalAcc
Lasso (RM+AMS)	0.66	0.82	0.74
Sigmoid cross-entropy	**0.79**	**0.86**	**0.82**
Sigmoid cross-entropy (nd)	0.76	0.85	0.80
Softmax	0.41	0.94	0.68
Softmax (nd)	0.42	0.93	0.68
One-against-all	0.42	0.94	0.68
One-against-all (nd)	0.41	0.95	0.68

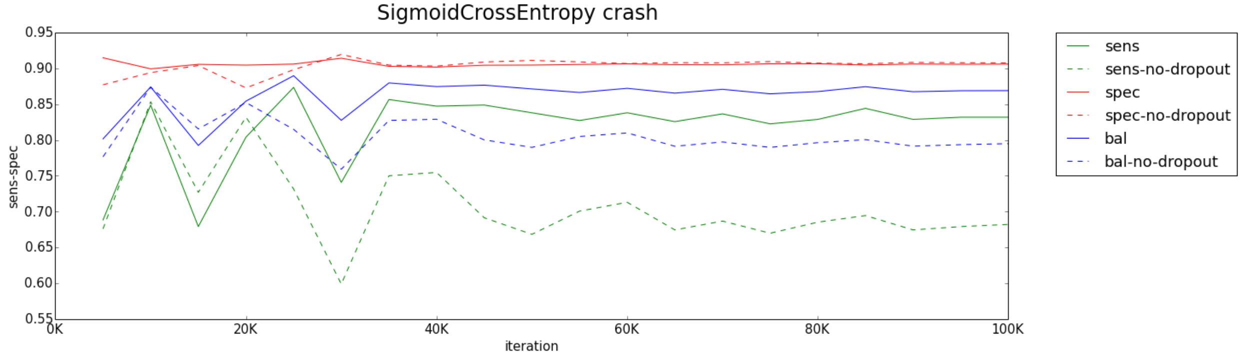


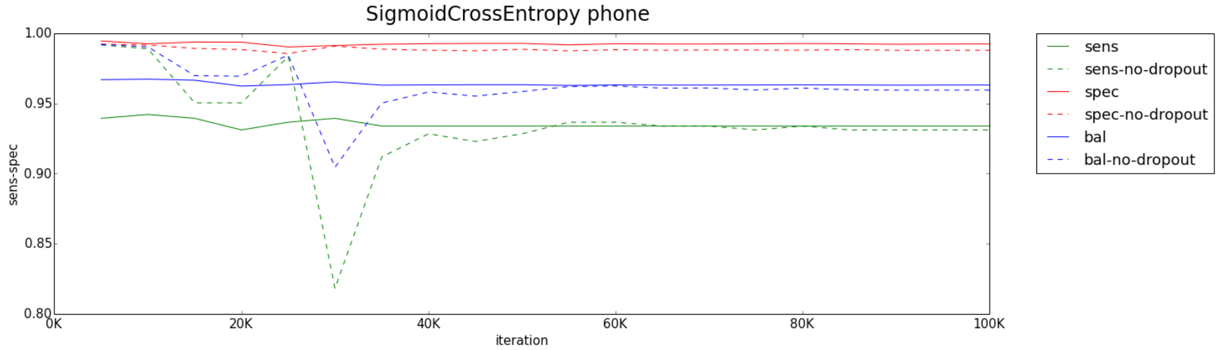


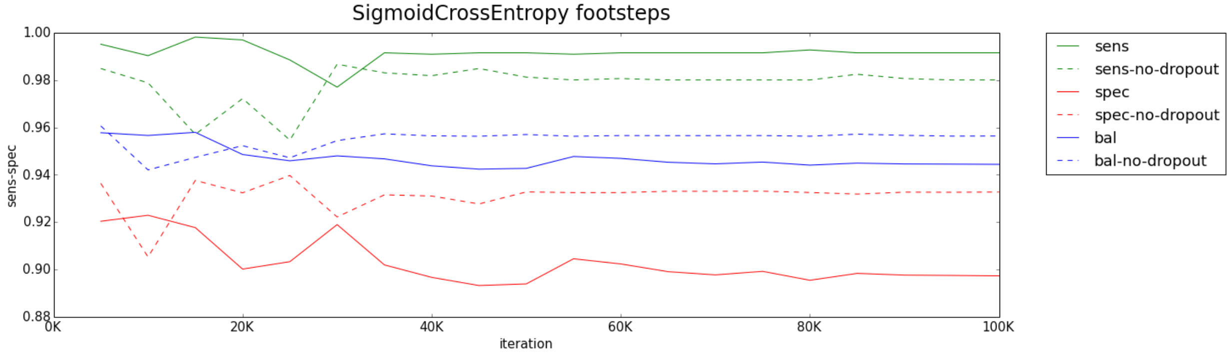

SigmoidCrossEntropy dog
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout

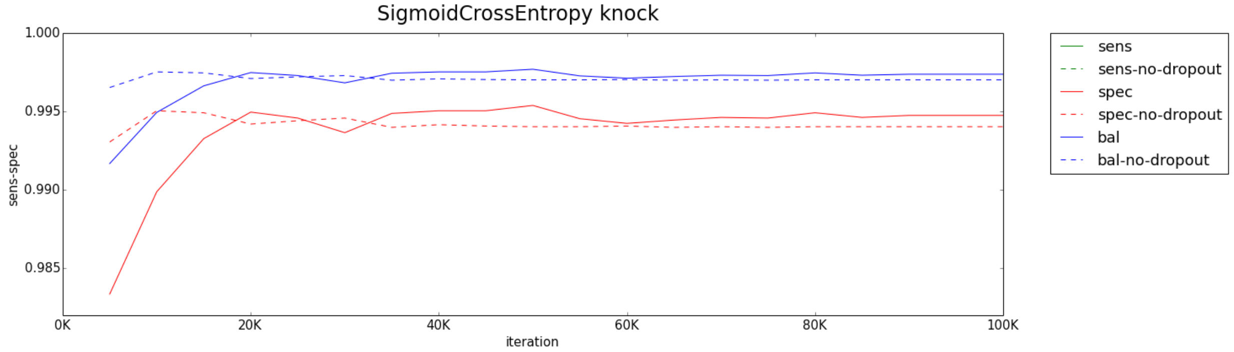

SigmoidCrossEntropy femaleSpeech
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout


SigmoidCrossEntropy fire
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout


SigmoidCrossEntropy crash
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout


SigmoidCrossEntropy phone
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout


SigmoidCrossEntropy footsteps
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout


SigmoidCrossEntropy knock
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout

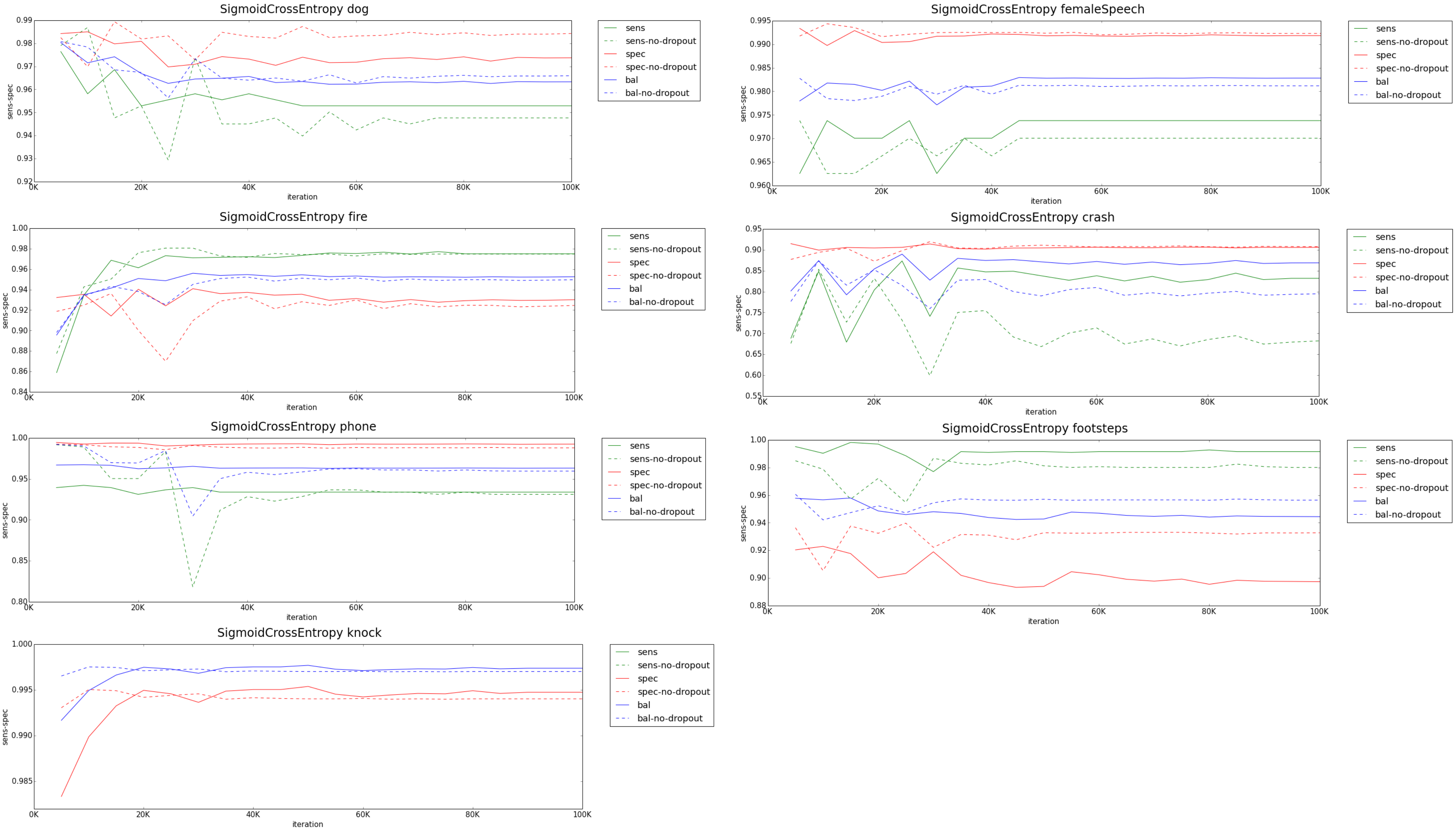
Conclusions

- Performance of Lasso on full feature set was in most cases better than or equal to Lasso on RM+AMS
- DNNs on raw ams and ratemaps achieved an overall similar performance compared to Lasso on L-Moments
  - Sometimes, poor balanced accuracy on test set due to low sensitivity
    - During training: sensitivity always higher than specificity
    - Maybe due to strong differences between sounds in training and test set for some classes
    - DNN might need more diverse sound examples to train on in order to generalize well for these classes, or stronger regularization
- The 3 different DNN architectures performed similarly
  - In terms of the stability against the sensitivity issue:
    sigmoid cross-entropy >> softmax >> one-vs-rest
- Effect of drop-out: ambivalent

Thank you.

Results from Lasso Model

	Ratemap + AMS features				Ratemap features				Full Feature Set
	Sensitvity	Specificity	BalAcc	Perf	Sensitvity	Specificity	BalAcc	Perf	Perf
female speech	0.996	0.998	0.997	0.997	0.977	0.983	0.980	0.980	1
alarm	0.950	0.873	0.911	0.903	0.919	0.831	0.875	0.867	0.92
baby	0.957	0.960	0.958	0.958	0.961	0.895	0.928	0.921	0.98
fire	0.950	0.945	0.947	0.947	0.893	0.855	0.874	0.873	0.94
knock	0.989	0.992	0.990	0.990	0.912	0.921	0.917	0.917	0.99
phone	0.992	0.942	0.967	0.958	0.994	0.890	0.942	0.922	0.94
dog	0.984	0.955	0.970	0.967	0.997	0.941	0.969	0.958	0.99
piano	0.830	0.929	0.878	0.868	0.633	0.947	0.790	0.738	0.97
footsteps	0.995	0.956	0.976	0.969	0.927	0.878	0.902	0.899	0.94
crash	0.910	0.842	0.876	0.871	0.777	0.667	0.722	0.717	0.88
engine	0.657	0.816	0.736	0.725	0.797	0.764	0.780	0.780	0.83



SigmoidCrossEntropy dog
SigmoidCrossEntropy femaleSpeech
SigmoidCrossEntropy fire
SigmoidCrossEntropy crash
SigmoidCrossEntropy phone
SigmoidCrossEntropy footsteps
SigmoidCrossEntropy knock
sens-spec
iteration
sens
sens-no-dropout
spec
spec-no-dropout
bal
bal-no-dropout