Music Technology Group
EXCELENCIA MARÍA DE MAEZTU
pandora®

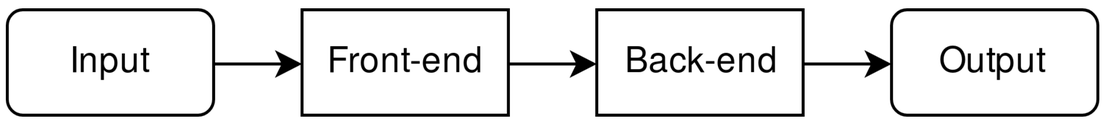
Jordi Pons, Oriol Nieto, Matt Prockup,
Erik Schmidt, Andreas Ehmann and Xavier Serra
contact: *jordipons.me - @jordiponsdotme*

End-to-end learning for music audio tagging at scale

waveform > spectrogram?

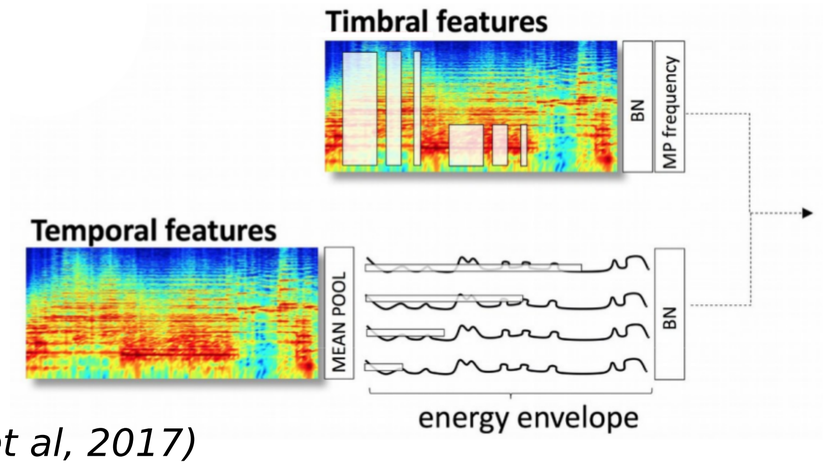
can waveform-based deep learning models achieve better performance than spectrogram-based ones?

YES! If enough training data is avialable



Spectrogram front-end

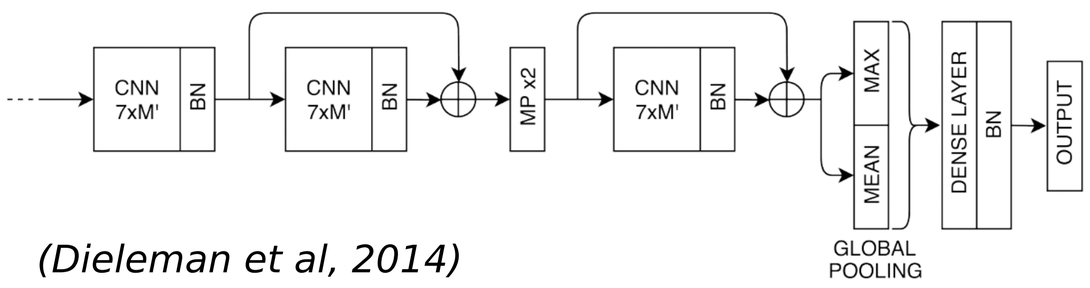
heavily based on domain knowledge



(Pons et al, 2017)

Shared back-end

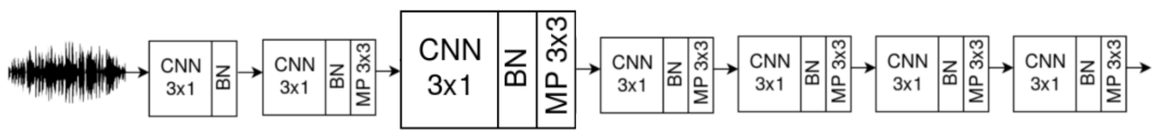
for a fair front-end comparison



(Dieleman et al, 2014)

Waveform front-end

an assumption-free model



(Lee et al, 2017)

MagnaTagATune dataset: 25k songs

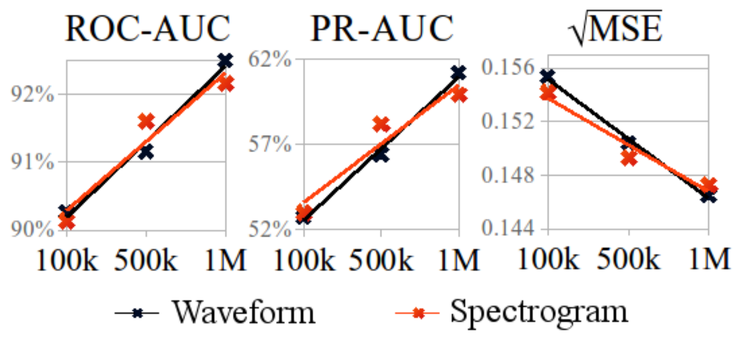
Waveform models	*ROC AUC*	*PR AUC*	*# param*
Dieleman et al.	85.58	29.59	194k
SampleCNN	88.56	34.38	2.4M
Waveform model (ours)	**89.05**	**34.92**	11.8M
Waveform model (smaller)	88.94	34.47	1.3M
Spectrogram models			
Timbre CNN	89.07	34.92	220k
VGG	89.99	37.56	450k
Spectrogram model (ours)	**90.40**	**38.11**	5M
Spectrogram model (smaller)	90.28	37.55	222k

few training data?

spectrogram > waveform

Private dataset: 1.2M songs

Models	*train size*	*ROC AUC*	*PR AUC*	$\sqrt{MSE}$
Baseline	1.2M	91.61%	54.27%	0.1569
Waveform	1M	**92.50%**	**61.20%**	**0.1465**
Spectrogram	1M	92.17%	59.92%	0.1473
Waveform	500k	91.16%	56.42%	0.1504
Spectrogram	500k	91.61%	58.18%	0.1493
Waveform	100k	90.27%	52.76%	0.1554
Spectrogram	100k	90.14%	52.67%	0.1542



lots of training data?

waveform > spectrogram

Million Song Dataset: 250k songs

Waveform models	*ROC AUC*	*PR AUC*	*# param*
SampleCNN	88.12	-	2.4M
SampleCNN multi-level & multi-scale	**88.42**	-	-
Waveform model (ours)	87.41	28.53	5.3M
Spectrogram models			
VGG + RNN	86.2	-	3M
Multi-level & multi-scale	**88.78**	-	-
Spectrogram model (ours)	**88.75**	**31.24**	5.9M

a reasonable ammount of training data?

spectrogram > waveform

Qualitative results

Bias towards predicting popular tags

"lead vocals", "English" or "male vocals"

Predicting each tag independently vs. predicting all tags together

"East Coast" vs. "West Coast"

"Baroque period" vs. "Classic period"