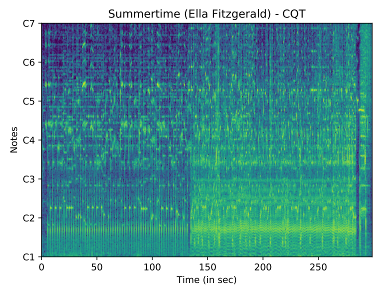
COMBINING MUSICAL FEATURES FOR COVER DETECTION

Guillaume Doras, Furkan Yesiler, Joan Serrà, Emilia Gómez, Geoffroy Peeters

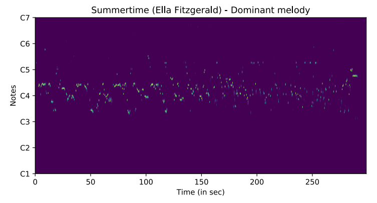
Recent studies addressed the automatic cover detection problem with the metric learning paradigm, using various input features:

full spectrum features

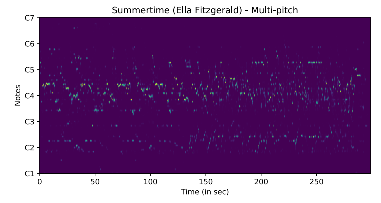


CQT [1]

melodic features

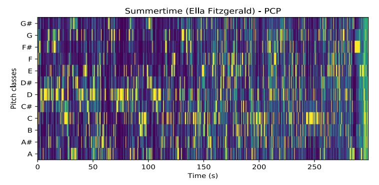


dominant melody [2]

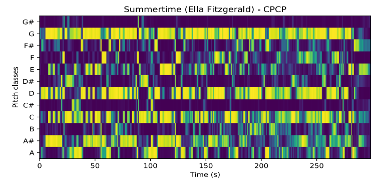


multi-pitch [3]

harmonic features

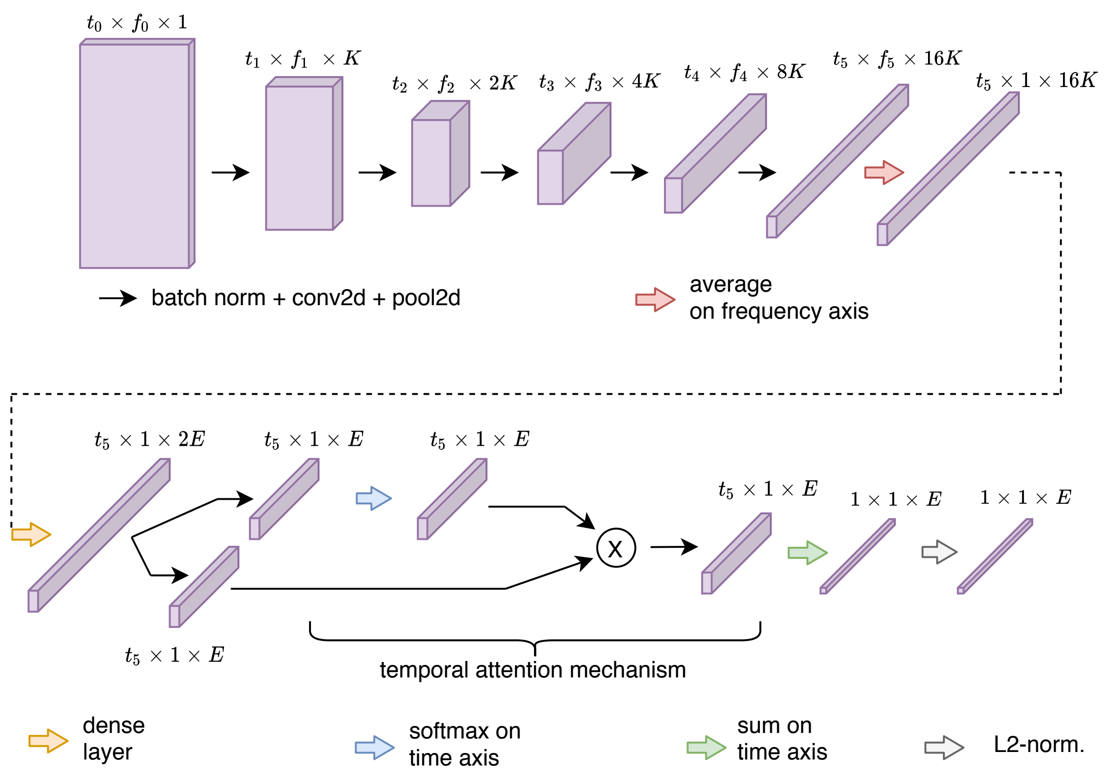


chroma [4]



Crema-PCP [5,6]

We compare these features with the same model using a time attention mechanism:

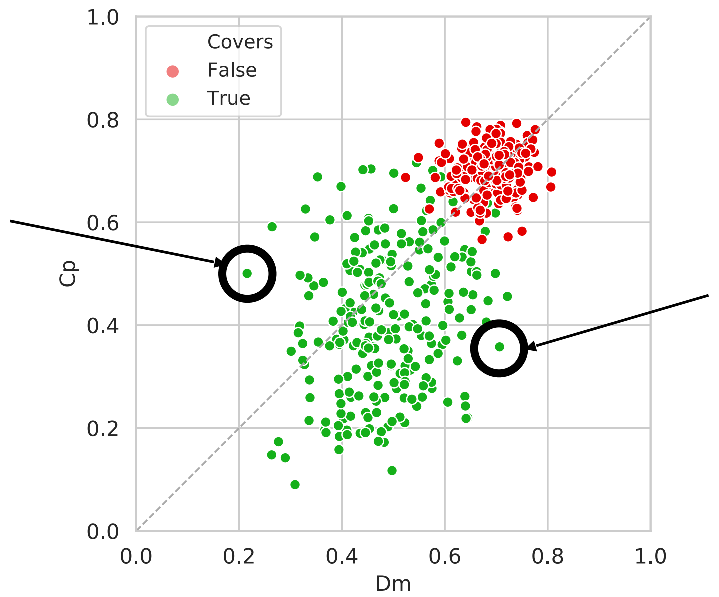

$t_0 \times f_0 \times 1$
$t_1 \times f_1 \times K$
$t_2 \times f_2 \times 2K$
$t_3 \times f_3 \times 4K$
$t_4 \times f_4 \times 8K$
$t_5 \times f_5 \times 16K$
$t_5 \times 1 \times 16K$
batch norm + conv2d + pool2d
average on frequency axis
$t_5 \times 1 \times 2E$
$t_5 \times 1 \times E$
$t_5 \times 1 \times E$
$t_5 \times 1 \times E$
$1 \times 1 \times E$
$1 \times 1 \times E$
$t_5 \times 1 \times E$
temporal attention mechanism
dense layer
softmax on time axis
sum on time axis
L2-norm.

➡ On two publicly available cover datasets, Crema-PCP — a harmonic feature — consistently yield the best results, followed by melodic features.

	Da-TACOS			SHS$_{4\text{-}}$		
Input	MAP	MT@10	MR1	MAP	MT@10	MR1
Cq	0.215	2.468	94	0.397	0.718	886
Dm	0.311	3.521	111	0.412	0.722	1431
Mp	0.293	3.290	**71**	0.422	0.760	**862**
Ch	0.121	1.476	117	0.174	0.371	1465
Cp	**0.375**	**4.084**	86	**0.499**	**0.842**	1169

(Cq = CQT, Dm = dominant melody, Mp = multi-pitch, Ch = chroma, Cp = CPCP)

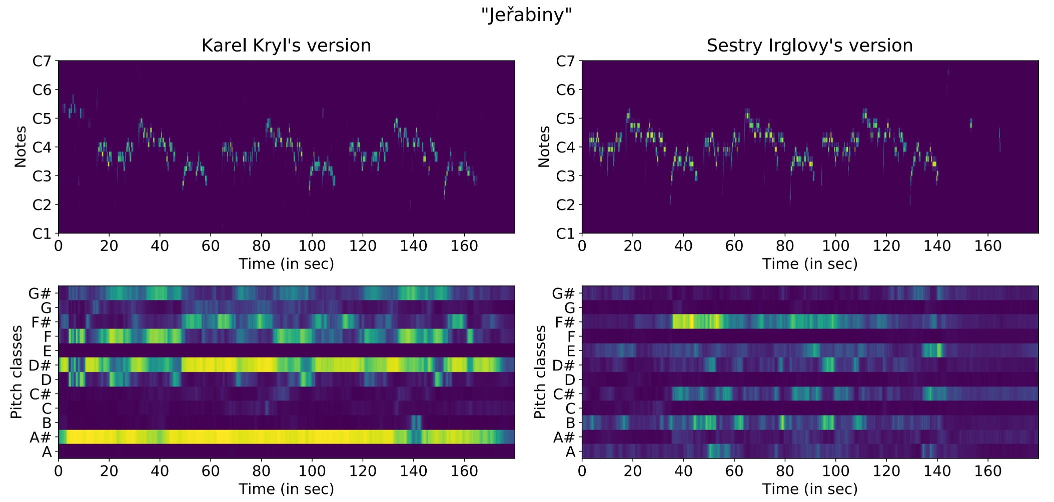
These features do not encode the same information:



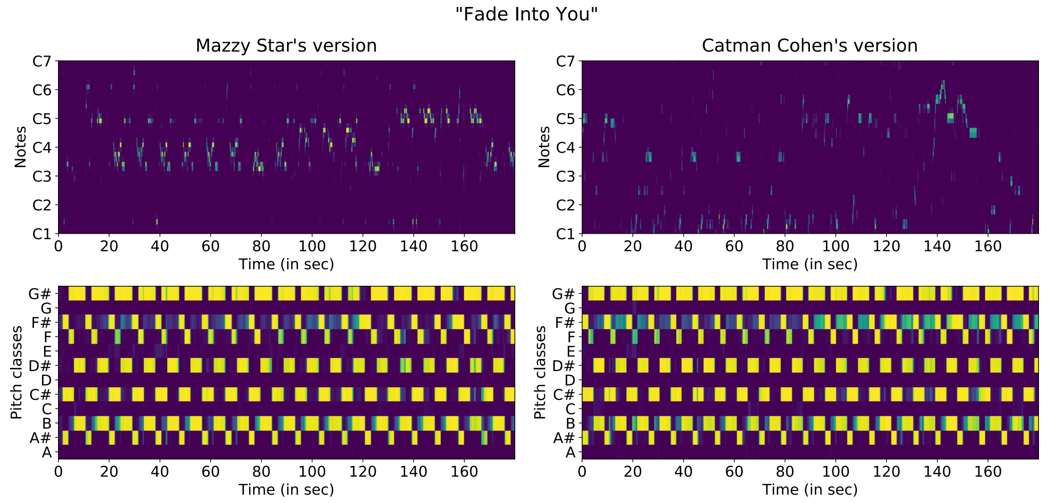
This cover pair obtains a high distance with the Crema-PCP (Cp), but not with the dominant melody (Dm).

This cover pair obtains a high distance with the dominant melody (Dm), but not with the Crema-PCP (Cp).

For instance, these two covers have a similar melody, but a different harmonic structure…



… while these two covers have a different melody, but a similar harmonic structure:



(see and listen to more examples on the Slack channel).

➡ This suggest that different features are complementary, and that merging them could benefit of this complementarity

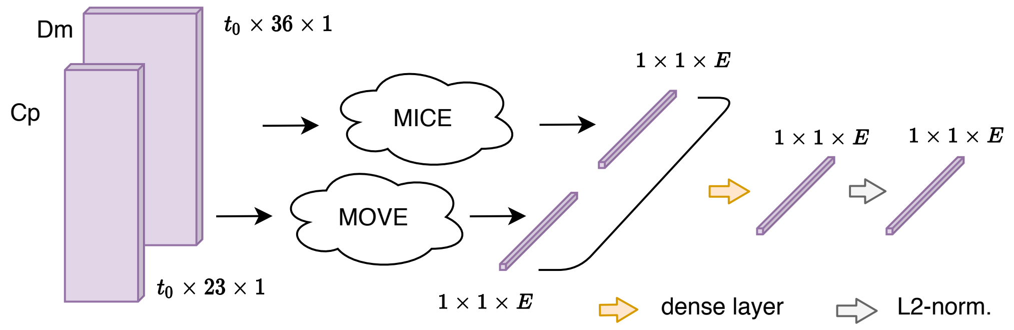
We combine these features with a simple averaging scheme — define each pair of songs (x, y) new distance as the average of their distances obtained for different features, e.g:

$$d(x,y) = \frac{d_{Dm}(x,y) + d_{Cp}(x,y)}{2}$$

➡ Combining dominant melody — a melodic feature and Crema-PCP — a harmonic feature — yields the best results.

Test set	Da-TACOS			SHS$_{4\text{-}}$		
Input	MAP	MT@10	MR1	MAP	MT@10	MR1
Cq+Dm	0.359	4.002	62	0.590	0.982	567
Cq+Mp	0.324	3.603	62	0.530	0.909	623
Cq+Cp	0.427	4.636	46	0.621	1.024	581
Dm+Mp	0.394	4.347	61	0.571	0.956	614
Dm+Cp	**0.547**	**5.861**	**37**	**0.679**	**1.098**	**529**
Mp+Cp	0.496	5.330	40	0.627	1.034	593

We also trained a new model to learn to combine these features:



➡ Combining musical features yields new SoA results:

	Da-TACOS			SHS$_{4-}$		
Input	MAP	MT@10	MR1	MAP	MT@10	MR1
Dm (MICE)	0.360	4.032	94	0.412	0.722	1431
Cp (MOVE)	0.484	5.214	59	0.533	0.890	1188
Dm+Cp (A)	0.621	6.613	32	**0.697**	**1.120**	**517**
Dm+Cp (LF-a)	0.570	6.101	**29**	0.617	1.017	686
Dm+Cp (LF-b)	0.592	6.318	32	0.655	1.059	655
Dm+Cp (LF-c)	**0.635**	**6.744**	30	0.660	1.080	657
Doras et al. [3]	n/a	n/a	n/a	0.323	0.615	1476
Yesiler et al. [6]	0.507	-	40	n/a	n/a	n/a

[1] Yu et al., "Learning a representation for cover song identification using convolutional neural network", ICASSP 2020
[2] Doras and Peeters, "Cover detection using dominant melody embeddings", ISMIR 2019
[3] Doras and Peeters, "A prototypical triplet loss for cover detection", ICASSP 2020
[4] Xu et al., "Key-invariant convolutional neural network toward efficient cover song identification", ICME 2019
[5] McFee and Bello, "Structured training for large-vocabulary chord recognition", ISMIR 2017
[6] Yesiler et al., "Accurate and scalable version identification using musically-motivated embeddings", ICASSP 2020