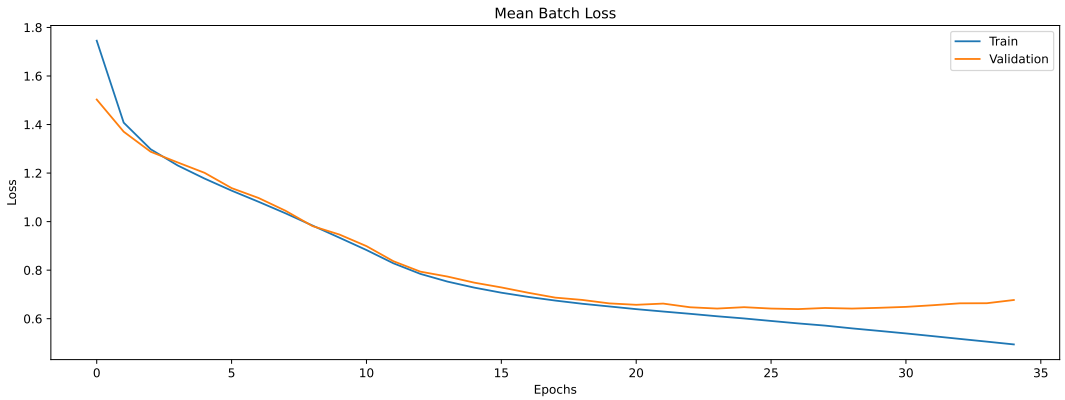


Fig. 1: Normalized temperature-scaled cross-entropy (NT-Xent) loss evolution over 35 epochs. The blue curve represents the training mean batch loss over 28,800 batches while the orange curve represents the validation mean batch loss over 6400 batches. As one can notice, the ZeroNS model starts to overfit after approximately 20 epochs. This motivated us to stop training.

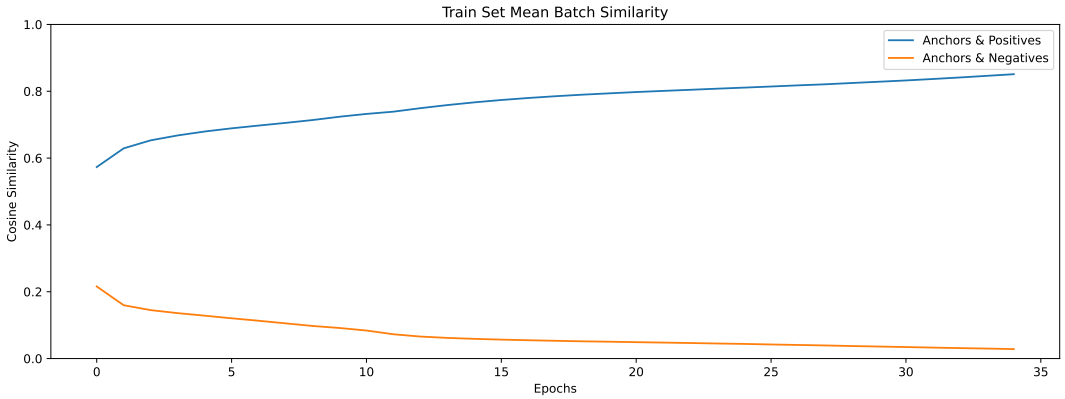


Fig. 2: Cosine similarity evolution on the training set over 35 epochs. The blue curve represents the mean cosine similarity between synchronized percussive parts (*positives*) and non-percussive parts (*anchors*). The orange curve represents the mean cosine similarity between non-synchronized percussive parts (*negatives*) and non-percussive parts (*anchors*).

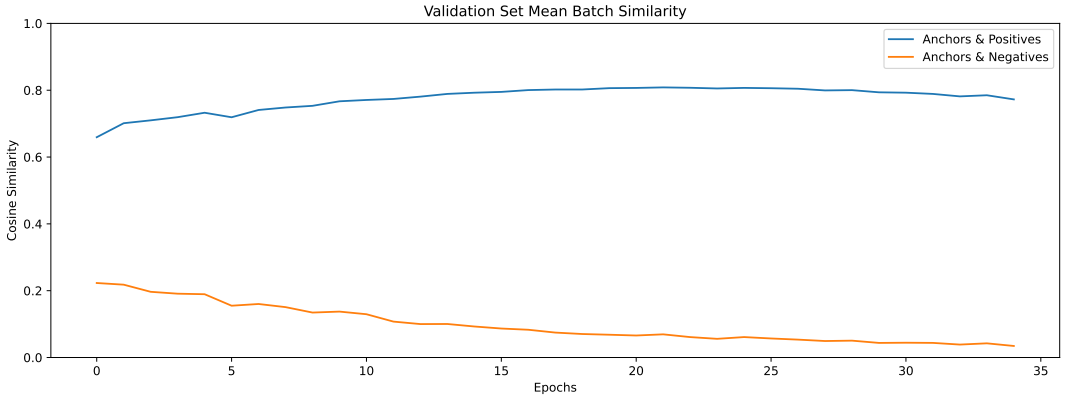
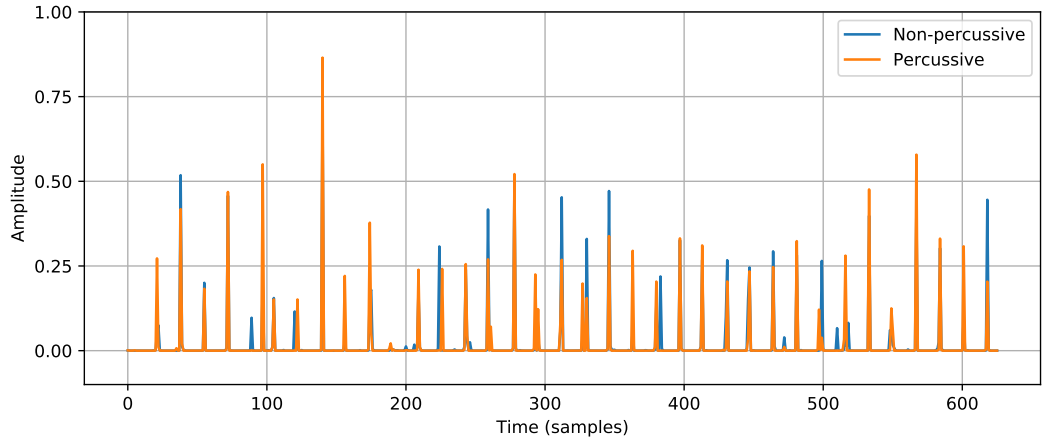
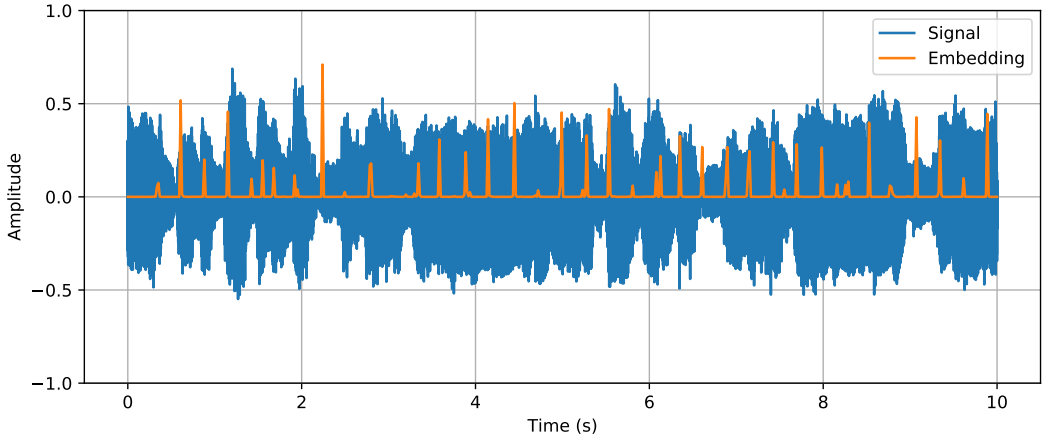


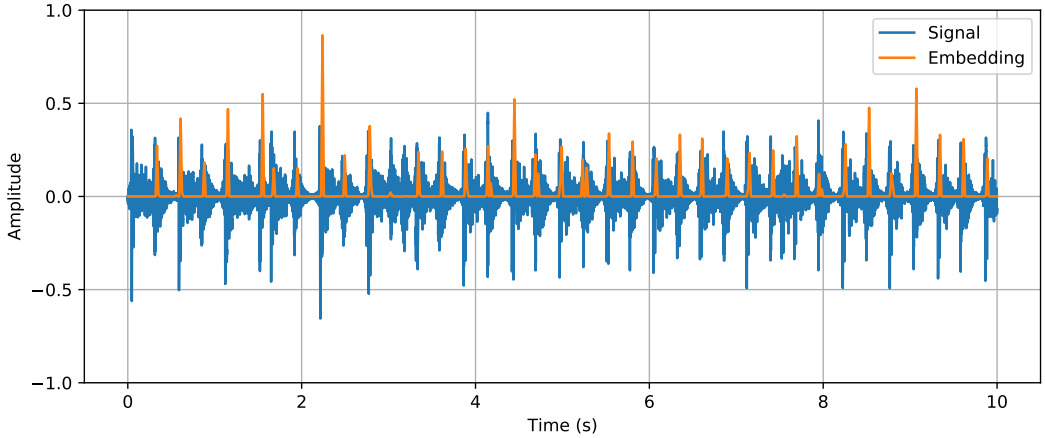
Fig. 3: Cosine similarity evolution on the validation set over 35 epochs. The blue curve and orange curves represent the same values as in the previous figure.



(a) Overlapped Embeddings

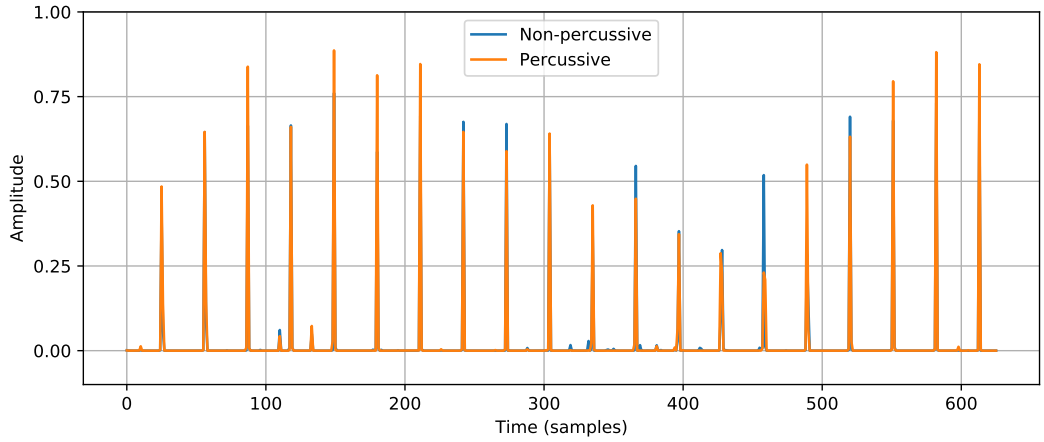


(b) Overlapped Non-percussive Waveform and Embedding

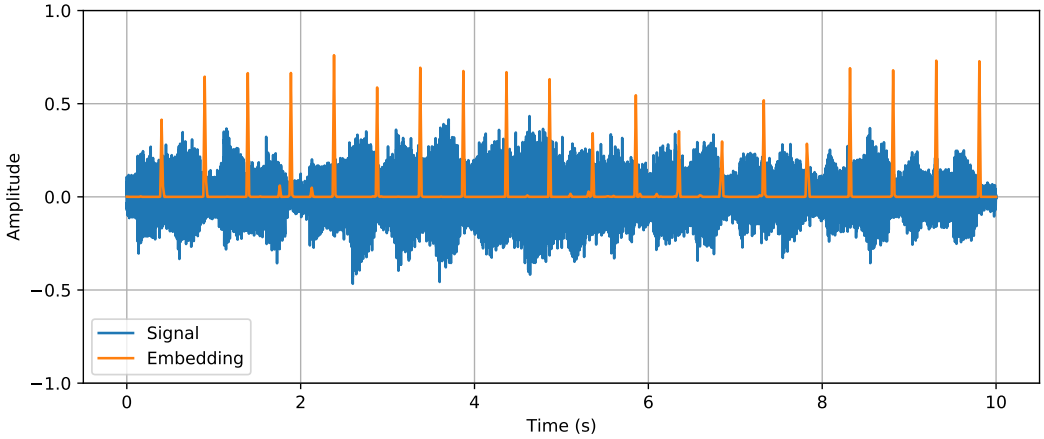


(c) Overlapped Percussive Waveform and Embedding

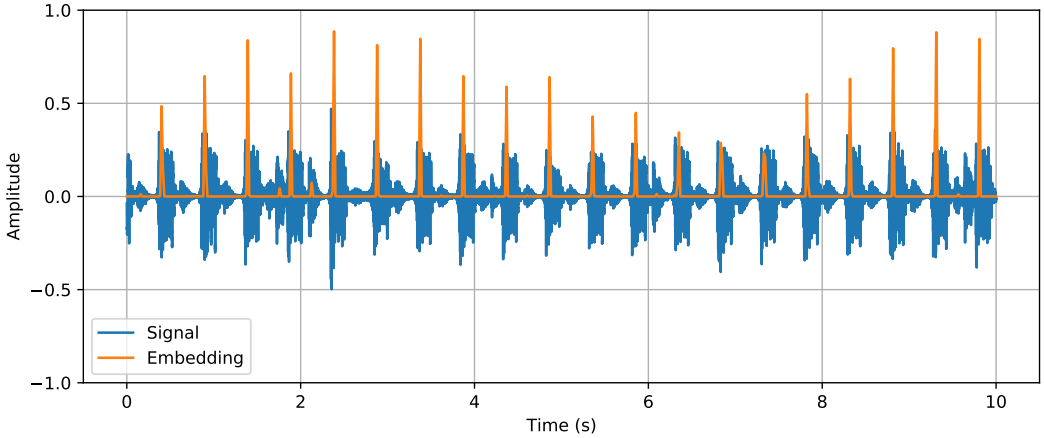
Fig. 4: First example signal. The cosine similarity between percussive and non-percussive embeddings is equal to 0.861 here.



(a) Overlapped Embeddings

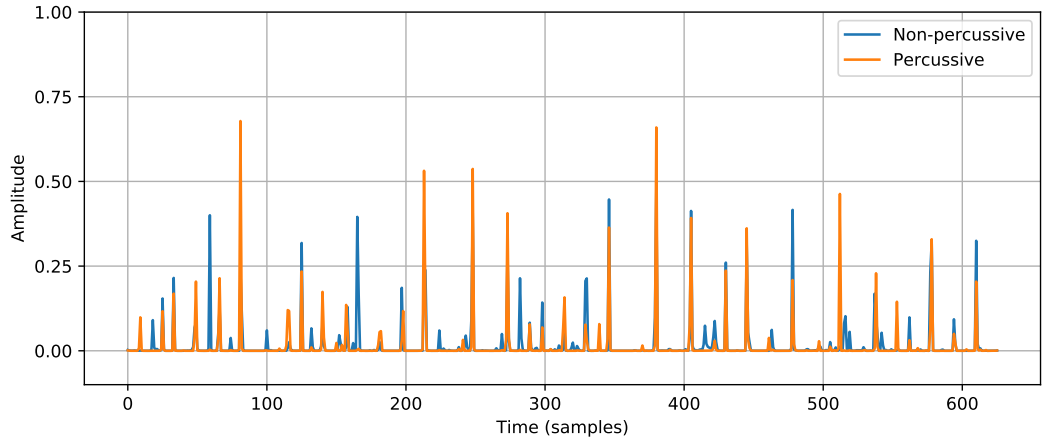


(b) Overlapped Non-percussive Waveform and Embedding

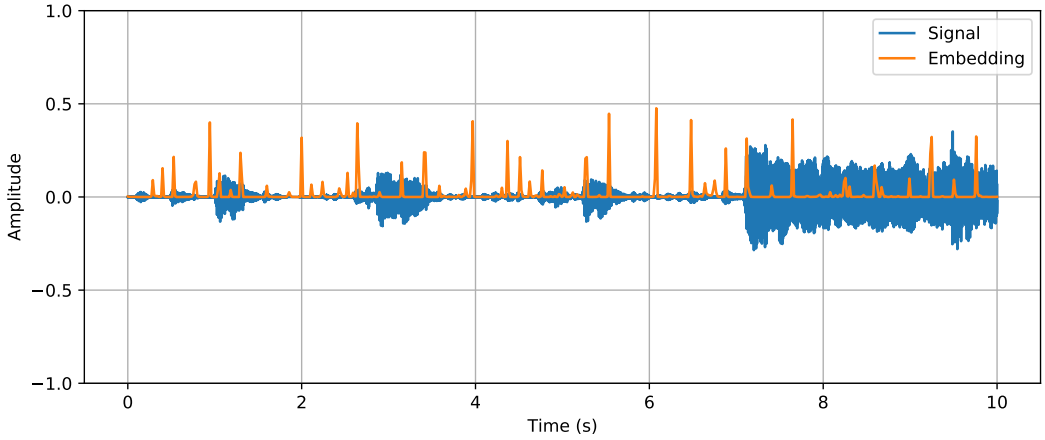


(c) Overlapped Percussive Waveform and Embedding

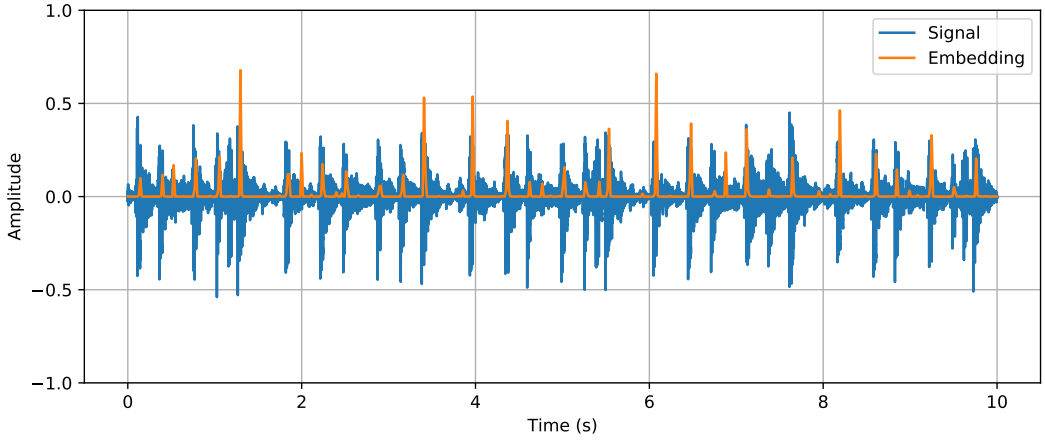
Fig. 5: Second example signal. The cosine similarity between percussive and non-percussive embeddings is equal to 0.971 here.



(a) Overlapped Embeddings



(b) Overlapped Non-percussive Waveform and Embedding



(c) Overlapped Percussive Waveform and Embedding

Fig. 6: Third example signal. The cosine similarity between percussive and non-percussive embeddings is equal to 0.747 here. Notice how the embedding during the "silent" parts of the non-percussive signal still contains peaks that could be seen as musical beats.