Objective Function

Our training data consists of a set $\mathcal{P}$, such that $(x,y) \in \mathcal{P}$ indicates that $x$ is a feature vector in some sequence from one modality which is aligned to $y$ in a matching sequence from another modality. We then construct $\mathcal{N}$ by repeatedly choosing two pairs $(x_1,y_1),(x_2,y_2) \in \mathcal{P}$ and swapping entries to construct $(x_1,y_2),(x_2,y_1) \in \mathcal{N}$. Motivated by [2], we use the following objective function:

$$
\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y) \in \mathcal{P}} \| f(x) - g(y) \|^2 + \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N}} \max(0,m-\| f(x) - g(y) \|^2)
$$

where $f$ and $g$ are learned nonlinear functions, $\alpha$ is a parameter to control the importance of separating dissimilar items, and $m$ is a target separation of dissimilar pairs.

Results

To evaluate, we trained our model on a collection of MIDI and audio recording pairs which were pre-aligned using DTW. We then computed hash sequences for every entry in the MSD and a hold-out set of 1,537 MIDI files for which we knew a priori the correct match. We measured performance as the percentage of MIDI files in this test set where the correct match in the MSD ranked below a certain threshold. In the unimodal setting, this approach ranked the correct entry in the top 1% (corresponding to 10,000 entries) 95.9% of the time; for multimodal sequences, performance degraded slightly to 92.8%.

References

