Pseudo-Supervised Training Improves Unsupervised Melody Segmentation

Stefan Lattner, Carlos Eduardo Cancino Chacón and Maarten Grachten
Austrian Research Institute for Artificial Intelligence
Freyung 6/6, 1010 Vienna, Austria
http://www.ofai.at/research/impml

Abstract

An important aspect of music perception in humans is the ability to segment streams of musical events into structural units such as motifs and phrases. A promising approach to the computational modeling of music segmentation employs the statistical and information-theoretic properties of musical data, based on the hypothesis that these properties can (at least partly) account for music segmentation in humans. Prior work has shown that in particular the information content of music events, as estimated from a generative probabilistic model of those events, is a good indicator for segment boundaries. In this paper we demonstrate that, remarkably, a substantial increase in segmentation accuracy can be obtained by not using information content estimates directly, but rather in a bootstrapping fashion. More specifically, we use information content estimates computed from a generative model of the data as a target for a feed-forward neural network that is trained to estimate the information content directly from the data. We hypothesize that the improved segmentation accuracy of this bootstrapping approach may be evidence that the generative model provides noisy estimates of the information content, which are smoothed by the feed-forward neural network, yielding more accurate information content estimates.

1 Introduction

A prominent theory about human perception and cognition states that 'chunking' is a key mechanism in human information processing [Gobet *et al.*, 2001]. By internally representing information in 'chunks'—meaningful constituents—humans are capable of interpreting information more efficiently than when information is processed in terms of lower level information units. A prominent example of chunking has been shown in the context of chess [Gobet and Simon, 1998], where increased skill level is associated with more efficient chunking of information about board configurations. Moreover, chunking is involved more generally in visual [McCollough and Vogel, 2007] and acoustic/speech processing [Baddeley, 1966] tasks. Just as in speech, perception in terms of meaningful constituents is a principal trait of music cognition. This is immanent in the ubiquitous notion of constituent structure in music theory.

The formation of chunks involves grouping and segmentation of information. To account for those phenomena in music perception, a prominent approach from music theory and cognitive psychology has been to apply perceptual grouping mechanisms, such as those suggested by Gestalt psychology. *Gestalt principles*, such as the laws of proximity, similarity, and closure, were first discussed in visual perception [Wertheimer, 1938], and have been successfully applied to auditory scene analysis [Bregman, 1990] and inspired theories of music perception [Meyer, 1956; Narmour, 1990; Lerdahl and Jackendoff, 1983]. Narmour's Implication-Realization theory [Narmour, 1990], for example, uses measures of pitch proximity and closure that offer insight into how listeners perceive the boundaries between musical phrases. This type of theory-driven approach has given rise to various rule-based computational models of segmentation. This class of models relies upon the specification of one or more principles according to which musical sequences are grouped.

An alternative account of grouping and segmentation is based on the intuition that the distribution, or statistical structure of the sensory information, has an important effect on how we perceive constituent structure. This idea has been explored for different areas, such as vision [Glicksohn and Cohen, 2011], speech [Brent, 1999], and melody perception [Pearce *et al.*, 2010b]. The key idea is that the sensory information that comprises a chunk is relatively constant, whereas the succession of chunks (which chunk follows which) is more variable. In information-theoretic terms, this implies that the *information content* (informally: unexpectedness) of events within a chunk is lower than that of events that mark chunk boundaries. As a side note on vocabulary: We will use the term *segment*, rather than *chunk* in the rest of this paper, to express that we take an agnostic stance toward the precise nature of constituents, and rather focus on their demarcation.

While Gestalt principles are sometimes rather abstractly defined laws, information theory has a certain potential to formally describe and quantify such perceptive phenomenons. The Gestalt idea of grouping based on "good form" (i.e. Prägnanz), for example, has an information theoretic coun-

terpart in the work of [von Helmholtz, 2005], where human vision is assumed to resolve ambiguous perceptive stimuli by preferring the most probable interpretation. In addition, it is intuitively clear that in most real-word scenarios, the uncertainty about expectations (i.e. the entropy) tends to increase with higher distances from observed events in any relevant dimension. Thus, while a direct link between the two paradigms is beyond dispute, the question remains which of it is more parsimonious and might have given rise for the other to emerge as a perceptual mechanism.

Prior work has shown that the information content of music events, as estimated from a generative probabilistic model of those events, is a good indicator for segment boundaries in melodies [Pearce *et al.*, 2010a]. In this paper we demonstrate that, remarkably, a substantial increase in segmentation accuracy can be obtained by not using information content estimates directly, but rather in a bootstrapping fashion. More specifically, we use information content estimates computed from a generative model of the data as a target for a feed-forward neural network (FFNN) that is trained to estimate the information content directly from the data.

In an experimental setup, we compare our method to other methods in an evaluation against human segment boundary annotations. Moreover, we offer an explanation for the improved accuracy by describing how our method can be regarded as employing a form of *entropy regularization* [Grandvalet and Bengio, 2004].

The structure of the paper is as follows. In Section 2, we discuss statistical models for melody segmentation, as well as related work regarding the pseudo-supervised regularization scheme. In Section 3, we describe how we estimate the conditional probability and information content of notes using a Restricted Boltzmann Machine (RBM), how notes are represented as input to the model, how an FFNN is used to predict information content, and how the information content is used to predict segment boundaries. Section 4 describes the experimental setup for evaluation of the model. The results are presented and discussed in Section 5, and conclusions and future work are presented in Section 6.

2 Related work

A notable information theory driven method for melodic segmentation is based on IDyOM, a class of variable order markov models for capturing the statistical structure of music [Pearce, 2005]. After training on musical data, IDyOM can produce a variety of information-theoretic quantities for a given musical context, such as *entropy*, expressing how confidently the model can predict the continuation of the context, and *information content*, expressing how unexpected the actual continuation of the context is, under the model. In particular the information content has been shown to be a good indicator of segment boundaries in monophonic melodies, using adaptive thresholding to predict segment boundaries when the information content (IC) of a note is high with respect to the IC values of its predecessors [Pearce *et al.*, 2010a].

Pearce et al. compare their probabilistic melody segmentation method, along with some other information theoretic models inspired by [Brent, 1999], to several knowledge based methods for melody segmentation, notably *Grouper* [Temperley, 2001], *LBDM* [Cambouropoulos, 2001], and several grouping rules that are part of the Generative Theory of Tonal Music [Lerdahl and Jackendoff, 1983], as formalized in [Frankland and Cohen, 2004]. The results of this comparison (that are partly reported in Section 5) show that IDyOM predicts segment boundaries much better than simpler information-theoretic methods, although not as accurately as *Grouper* and *LBDM*.

In prior work, we have proposed a probabilistic segmentation method analogous to IDyOM, but using an RBM as a probabilistic model of the data, rather than variable order markov models [Lattner *et al.*, 2015]. This method was shown to predict segment boundaries less accurately than *Grouper* and *LBDM*, but better than IDyOM. In Sections 3.1 to 3.3, we recapitulate the RBM based approach to compute IC values for melody notes. The actual contribution of this paper is an extension of this approach to what we call a *pseudo-supervised* scenario (Section 3.4). In this setting, rather than using the IC estimations from the RBM directly for predicting segments, they are used as targets for a FFNN, which is trained to predict IC values directly from the data (without a probabilistic model).

Although the term pseudo-supervised does not (yet) seem to have a well-established meaning, our use of the term is compatible with its use in [Nøklestad, 2009], in the sense that a supervised approach is used to predict targets that are computed from the input data, rather than relying on hand-labeled (or otherwise authoritative) targets. The automatically generated targets (in this case IC values) are not themselves the actual targets of interest (the boundary segments), but are instrumental to the prediction of the actual targets.

Similar methods are proposed by [Lee, 2013] and [Hinton *et al.*, 2014], where supervised models are used to generate targets (pseudo labels or soft-targets) from new data. But in contrast to a *pseudo-supervised* approach, these methods require hand-labeled data, and are strictly taken a *semi-supervised* approach, in which predictive models are trained partly in an unsupervised manner, and partly using hand-labeled data.

In general, both semi- and pseudo-supervised learning approaches benefit from the use of unlabeled information by using Bayesian approaches, which make assumptions over the distribution of unlabeled data. From a formal standpoint, these techniques act as regularizers of the model parameters, and thus, prevent overfitting. Approaches like *entropy regularization* [Grandvalet and Bengio, 2004] use the principle of maximum entropy to select a prior distribution of the model parameters, and then optimize the model in Maximum a Posteriori (MAP) sense.

3 Method

In this Section, we describe the methods used to predict melody segment boundaries. We start by describing how conditional probabilities of music events can be estimated by training an RBM as a probabilistic model of the data (Section 3.1). The representation of music events is described in Section 3.2. Section 3.3 details how the IC of music events

is computed based on their estimated conditional probabilities. In Section 3.4, we describe how training a supervised model using IC values as (pseudo) targets can act as a form of regularization. Finally, Section 3.5 describes how segment boundaries are predicted from sequences of IC values.

3.1 Probability approximation through Monte Carlo techniques

An RBM is a stochastic Neural Network with two layers, a visible layer with units $\mathbf{v} \in \{0,1\}^r$ and a hidden layer with units $\mathbf{h} \in \{0,1\}^q$ [Hinton, 2002]. The units of both layers are fully interconnected with weights $\mathbf{W} \in \mathbb{R}^{r \times q}$, while there are no connections between the units within a layer.

In a trained RBM, the marginal probability distribution of a visible configuration $\mathbf{v}$ is given by the equation

$$p(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}, \tag{1}$$

where $E(\mathbf{v}, \mathbf{h})$ is an energy function. The computation of this probability distribution is intractable, because it requires summing over all possible joint configurations of $\mathbf{v}$ and $\mathbf{h}$ as

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}. \tag{2}$$

However, with Monte Carlo techniques it is possible to approximate the probability of a visible unit configuration $\mathbf{v}$. To that end, for N randomly initialized *fantasy particles*[1] $\mathbf{Q}$, we execute Gibbs sampling until thermal equilibrium. In the visible *activation vector* $\mathbf{q}_i$ of a fantasy particle i, element q_{ij} specifies the probability that visible unit j is on. Since all visible units are independent given $\mathbf{h}$, a single estimate based on one fantasy particles visible activation is computed as

$$p(\mathbf{v}|\mathbf{q}_i) = \prod_j p(v_j|q_{ij}). \tag{3}$$

As we are using binary units, such an estimate can be calculated by using a binomial distribution with one trial per unit. We average the results over N fantasy particles, leading to an increasingly close approximation of the true probability of $\mathbf{v}$ as N increases:

$$p(\mathbf{v}|\mathbf{Q}) = \frac{1}{N} \sum_{i}^{N} \prod_j \binom{1}{v_j} q_{ij}^{v_j} (1 - q_{ij})^{1-v_j}. \tag{4}$$

Posterior probabilities of visible units

When the visible layer consists of many units, N will need to be very large to obtain good probability estimates with the method described above. However, for conditioning a small subset of visible units $\mathbf{v}_s \subset \mathbf{v}$ on the remaining visible units $\mathbf{v}_c = \mathbf{v} \setminus \mathbf{v}_s$, the above method is very useful. This can be done by Gibbs sampling after randomly initializing the units $\mathbf{v}_s$ while clamping all other units $\mathbf{v}_c$ according to their initial state in $\mathbf{v}$. In Eq. 4, all $\mathbf{v}_c$ contribute a probability of 1, which results in the conditional probability of $\mathbf{v}_s$ given $\mathbf{v}_c$.

[1] See [Tieleman, 2008]

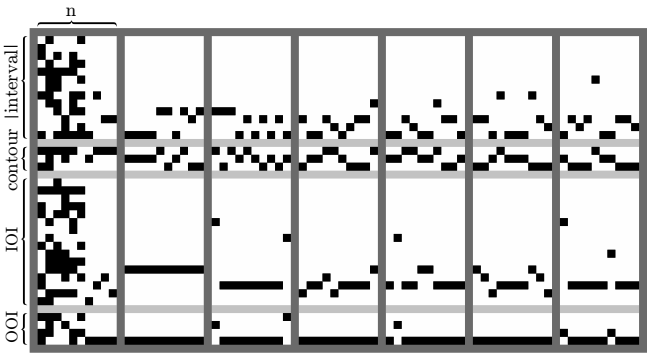

Figure 1: Seven examples of n-gram training instances (n=10) used as input to the RBM. Within each instance (delimited by a dark gray border), each of the 10 columns represents a note. Each column consists of four *one-hot* encoded viewpoints: |*interval*|, *contour*, *IOI* and *OOI* (indicated by the braces on the left). The viewpoints are separated by horizontal light gray lines for clarity. The first instance shows an example of noise padding (in the first six columns) to indicate the beginning of a melody.

We use this approach to condition the units belonging to the last time step of an n-gram on the units belonging to preceding time steps. For the experiments reported in this paper, we found that it is sufficient to use 150 fantasy particles and for each to perform 150 Gibbs sampling steps.

Training

We train a single RBM using *persistent contrastive divergence* (PCD) [Tieleman, 2008] with *fast weights* [Tieleman and Hinton, 2009], a variation of the standard *contrastive divergence* algorithm [Hinton *et al.*, 2006]. This method yields models which are well-suited for sampling, because it results in a better approximation of the likelihood gradient.

Based on properties of neural coding, sparsity and selectivity can be used as constraints for the optimization of the training algorithm [Goh *et al.*, 2010]. Sparsity encourages competition between hidden units, and selectivity prevents overdominance by any individual unit. A parameter μ specifies the desired degree of sparsity and selectivity, whereas another parameter ϕ determines how strongly the sparsity/selectivity constraints are enforced.

3.2 Data Representation

From the monophonic melodies, we construct a set of n-grams by using a sliding window of size n and a step size of 1. For each note in the n-gram, four basic features are computed: 1) absolute values of the pitch interval between the note and its predecessor (in semitones); 2) the contour (up, down, or equal); 3) inter-onset-interval; and 4) onset-to-offset-interval. The IOI and OOI values are quantized into semiquaver and quaver, respectively. Each of these four features is represented as a binary vector and its respective value for any note is encoded in a one-hot representation. The first n-1 n-grams in a melody are noise-padded to account for the first n-1 prefixes of the melody. Some examples of binary representations of n-grams are given in Figure 1).

3.3 Information Content

After we trained the model as described in Section 3.1, we estimate the probability of the last note conditioned on its preceding notes for each n-gram as introduced in Section 3.1. From the probabilities $p(e_t \mid e_{t-n+1}^{t-1})$ computed in this way, we calculate the IC as

$$h(e_t \mid e_{t-n+1}^{t-1}) = log_2 \frac{1}{p(e_t \mid e_{t-n+1}^{t-1})}, \tag{5}$$

where e_t is a note event at time step t, and e_k^l is a note sequence from position k to l of a melody. IC is a measure of the unexpectedness of an event given its context. According to a hypothesis of [Pearce *et al.*, 2010a], segmentation in auditory perception is determined by perceptual expectations for auditory events. In this sense, the IC relates directly to this perceived boundary strength, thus we refer to the IC over a note sequence as the *boundary strength profile* (BSP).

3.4 Pseudo-Supervised optimization

Algorithm 1: Pseudo-supervised training

Data: Set of n-grams : $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$

1 Train an RBM by optimizing the model parameters as

$$\tilde{\theta} = \operatorname*{argmax}_{\theta} \log p(\mathbf{v} \mid \theta) \tag{6}$$

2 Compute the set of *pseudo-targets* $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$ as

$$\mathbf{t}_t(\mathbf{v}_t; \tilde{\theta}) = h(e_t \mid e_{t-n+1}^{t-1}), \tag{7}$$

where $\mathbf{v}_t$ is the encoding of the n-gram $\{e_{t-n+1}, \ldots, e_t\}$, and $h(e_t \mid e_{t-n+1}^{t-1})$ is the IC computed as in Eq. (5).

3 Build a three layered FFNN and optimize it in a supervised way, using the set of pseudo-targets $\mathbf{T}$ as

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} \|\mathbf{t}(\mathbf{v}_t; \tilde{\theta}) - \mathbf{y}(\mathbf{v}_t; \theta)\|^2, \tag{8}$$

where $\mathbf{y}(\mathbf{v}_t; \theta)$ is the output of the FFNN for $\mathbf{v}_t$ given the model parameters θ.

4 **return** Model parameters $\hat{\theta}$

In contrast to our prior work, we do not use the BSP estimated from the RBM for segmentation. Instead, we train an FFNN to predict the estimated BSP directly from the data in a non-probabilistic manner, and use that curve for predicting segment boundaries (by the procedure described in Section 3.5). This is a way of context sensitive smoothing, which is achieved by the generalization ability of the NN. Note that no labeled data is used at any stage of the processing pipeline. The fact that this approach still improves the segmentation results is evidence that the generative model, as described in Section 3.1, provides noisy IC estimates. This is either due to poor approximations to the actual IC by the model itself, or the data can be considered to be noisy with respect to prototypical segment endings.

The proposed pseudo-supervised training method is shown in Algorithm 1. Formally, this method is an approximate MAP estimation of the parameters using entropy regularization [Grandvalet and Bengio, 2004]. In this method, the MAP estimate of the model parameters is computed as

$$\theta_{MAP} = \operatorname*{argmax}_{\theta} \log p(\mathbf{v} \mid \theta) - \lambda H(\mathbf{t} \mid \mathbf{v}; \theta), \tag{9}$$

where $H(\mathbf{t} \mid \mathbf{v})$ is the conditional Shannon entropy of the targets given the inputs, and λ is a Lagrange multiplier. In the proposed algorithm, this approximation is obtained by independently optimizing $\log p(\mathbf{v} \mid \theta)$ (see Eq. (6)), and then minimizing Eq. (8), which is equivalent to maximizing

$$p(\mathbf{t} \mid \mathbf{v}; \theta, \beta) = \mathcal{N}(\mathbf{t} \mid \mathbf{y}(\mathbf{v}, \theta), \beta^{-1}), \tag{10}$$

where β is the precision (inverse variance) of the distribution. This precision can be found by minimizing the negative log-likelihood of the above probability to give

$$\beta = \frac{N}{\sum_i \|\mathbf{t}_i - \mathbf{y}(\mathbf{v}_i, \theta)\|^2}. \tag{11}$$

The Shannon entropy for this distribution is given by

$$\begin{aligned} H(\mathbf{t} \mid \mathbf{v}; \theta, \beta) &= \mathbb{E}\{-\log p(\mathbf{t} \mid \mathbf{v})\} \\ &= \frac{1}{2} \log\left(\frac{2\pi}{\beta}\right) + \frac{1}{2}, \end{aligned} \tag{12}$$

which is minimal, since $\sum_i \|\mathbf{t}_i - \mathbf{y}(\mathbf{v}_i, \theta)\|^2$ is minimal. Therefore, optimizing Eq. (8) is equivalent to minimizing the entropy term in Eq. (9).

We use the fact that the RBM is a generative model, and therefore, the pseudo targets $\mathbf{t}$ come from the computation of the IC from a probabilistically sound estimate of the input data. In this way, pseudo-supervised learning can be understood as a suboptimal entropy-regularized MAP model of the model parameters.

Training

To compute $\hat{\theta}$ in Equation (8), we use a three layered FFNN with sigmoid hidden units and a single linear unit in the output layer. We pre-train the hidden layer with PCD and fine-tune the whole stack with Backpropagation, by minimizing the mean square error. As targets, we use the boundary strength values, estimated by the initial model described in Section 3.1.

After training, the outputs $\mathbf{y}(\mathbf{v}_i; \hat{\theta})$ of the FFNN are used as (improved) estimates of the information content of e_t, given $\{e_{i-n+1}, \ldots, e_{t-1}\}$.

3.5 Peak Picking

Based on the BSP described in Section 3.3 and the outputs $\mathbf{y}(\mathbf{v}_i; \hat{\theta})$ of the FFNN (see Algorithm 1), respectively, we need to find a discrete binary segmentation vector. For that, we use the peak picking method described in [Pearce *et al.*, 2010a]. This method finds all peaks in the profile and keeps those which are k times the standard deviation greater than the mean boundary strength, linearly weighted from the beginning of the melody to the preceding value:

$$S_n > k\sqrt{\frac{\sum_{i=1}^{n-1}(w_i S_i - \bar{S}_{w,1\ldots n-1})^2}{\sum_1^{n-1} w_i}} + \frac{\sum_{i=1}^{n-1} w_i S_i}{\sum_1^{n-1} w_i}, \quad (13)$$

where S_m is the m-th value of the BSP, and w_i are the weights which emphasize recent values over those of the beginning of the song (triangular window), and k has to be found empirically.

4 Experiment

To allow for a comparison of our segmentation method with other approaches, the experimental setup of the proposed method follows [Pearce *et al.*, 2010a] both in terms of data and procedure.

4.1 Data

For training and testing, we use the Essen Folk Song Collection (EFSC) [Schaffrath, 1995], a widely used corpus in music information retrieval (MIR). This database consists of more than 6000 transcriptions of folksongs primarily from Germany and other European regions. Due to the fact that phrase markers are encoded, the EFSC is one of the most used collections for testing computational models of music segmentation.

In accordance with [Pearce *et al.*, 2010a], we use the *Erk* subset of the EFSC, which consists of 1705 German folk melodies with a total of $78,995$ note events. Encoded phrase boundary annotations in the corpus constitute the baseline of about 12% positive examples.

4.2 Procedure

The model is trained and tested on the data described in Section 4.1 with various n-gram lengths between 3 and 10. For each n-gram length, we perform 5-fold cross-validation and average the results over all folds. Similar to the approach in [Pearce *et al.*, 2010a], after computing the BSPs, we evaluate different k from the set $\{0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00\}$ (initial IC estimation), and $\{0.24, 0.26, 0.28, 0.30, 0.32, 0.34, 0.36\}$ (after pseudo-supervised optimization), and choose the value that maximizes F1 for the respective n-gram length. To make results comparable to those reported in [Pearce *et al.*, 2010a], the output of the model is appended with an implicit (and correct) phrase boundary at the end of each melody.

Since the hyper-parameters of the model are interdependent, it is infeasible to exhaustively search for the optimal parameter setting. We have manually chosen a set of hyper-parameters that give reasonable results for the different models tested. For the initial IC estimation, we use 200 hidden units, a momentum of 0.6, and a learning rate of 0.0085 which we linearly decrease to zero during training. With increasing n-gram length we linearly adapt the batch size from 250 to 1000. In addition, we use 50% dropout on the hidden layer and 20% dropout on the visible layer.

The fast weights used in the training algorithm (see Section 3.1) help the fantasy particles mix well, even with small learning rates. The learning rate of the fast weights is increased from 0.002 to 0.007 during training. The training is continued until convergence of the parameters (typically between 100 and 300 epochs). The sparsity parameters (see Section 3.1) are set to $\mu = 0.04$, and $\phi = 0.65$, respectively. In addition, we use a value of 0.0035 for $L2$ weight regularization, which penalizes large weight coefficients.

For pre-training of the first layer in the FFNN, we change the learning rate to 0.005, leave the batch size constant at 250 and increase the weight regularization to 0.01. We again use dropout, for both the pre-training and the fine-tuning.

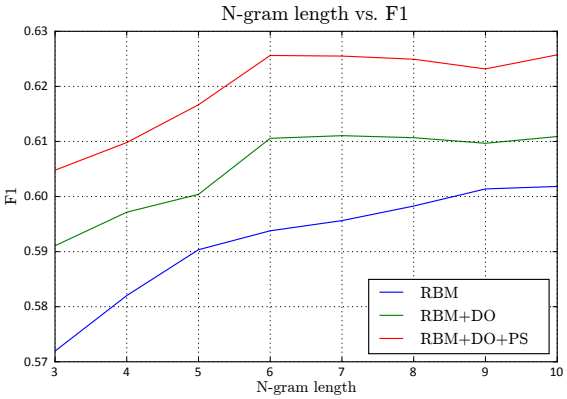

Figure 2: F1 scores for different N-gram lengths and methods.

5 Results and discussion

Figure 2 shows the F1 scores for different N-gram lengths and methods. By using dropout, the F1 score increases considerably, as dropout improves the generalization abilities of the RBM. With the pseudo-supervised approach, again a significant improvement of the classification accuracy can be achieved. This is remarkable, considering that no additional information was given to the FFNN, the improvement was based solely on context-sensitive smoothing.

Figure 3 shows the adaptation of single IC values through the pseudo-supervised optimization. Some previously true positives are erroneously regularized downwards (green lines from upper left to lower right), while some previously false negatives are correctly moved upwards (green lines from lower left to upper right). Quantitative tests show that our method increases IC values at boundaries more often than it decreases them. In general, if the initial BSP curve is correct in most cases, in pseudo-supervised training such regularities are detected and utilized.

Table 1 shows prediction accuracies in terms of precision, recall, and F1 score, both for our method and for the various alternative approaches mentioned in Section 2. The table shows that with the proposed method (RBM10+DO+PS), an information-theoretic approach is now on a par with a Gestalt-based approach (LBDM), while Grouper still provides the best estimates of melodic segment boundaries. However, Grouper exploits additional domain knowledge like musical parallelism, whereas the LBDM model, as well as (RBM10+DO+PS), are pure representatives of the Gestalt-based paradigm and the information-theoretic paradigm, re-

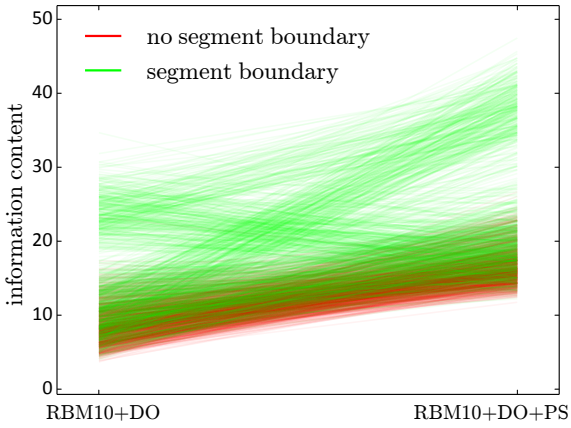

Figure 3: The effect of pseudo-training on estimated IC values; Line segments connect IC values estimated directly from the probabilistic model (RBM10+DO) with the corresponding IC values after pseudo-training (RBM10+DO+PS); Green lines indicate music events that mark a segment boundary, red lines indicate those that do not.

Model	Precision	Recall	F1
Grouper	0.71	0.62	0.66
LBDM	0.70	0.60	0.63
RBM10+DO+PS	**0.80**	**0.55**	**0.63**
RBM10+DO	0.78	0.53	0.61
RBM10	0.83	0.50	0.60
IDyOM	0.76	0.50	0.58
GPR 2a	0.99	0.45	0.58

Table 1: Results of the model comparison, ordered by F1 score. RBM results are shown for 10-grams of the initial RBM, the RBM with Dropout (DO) and the RBM with Pseudo-Supervised training (PS). Table adapted from [Pearce *et al.*, 2010a], with permission.

spectively.

The *GPR 2a* method is a simple rule that predicts a boundary whenever a rest occurs between two successive notes. Note how *GPR 2a* accounts for a large portion of the segment boundaries (approx. 45%). This implies that the challenge is mainly in recognizing boundaries that do not co-occur with a rest. For boundaries without rests, the pseudo-supervised approach yields an improvement of 3.7% in the F-score, while boundaries indicated by a rest did not improve any more (as for those boundaries the initial approach already yields an F-score of 0.99).

6 Conclusion and future work

In this paper, we show how a technique we call *pseudo-supervised* training improves the prediction accuracy of a probabilistic method for melody segmentation. Our method is a purely probabilistic method, that does not rely on any knowledge about musical structure. We use the information content (IC) of musical events (estimated from a probabilistic model) as a proxy for the actual target to be predicted (segment boundaries). With these *pseudo targets*, we train a feed-forward neural network. We show that segment boundaries estimated from the output of this network are more accurate than boundaries estimated from the *pseudo targets* themselves.

In this paper, we used the IC as estimated from an RBM, but the pseudo-supervised approach may benefit from including IC estimates from other models, such as IDyOM [Pearce, 2005]. In addition, there are other probabilistic architectures, such as conditional RBMs [Taylor *et al.*, 2006], that seem appropriate for estimating IC values from data. Furthermore, although the focus of this paper has been on IC, it is intuitively clear that IC is not the only factor that determines the perception of segment boundaries in melodies. Future experimentation is necessary to determine whether (combinations of) other information-theoretic quantities are also helpful in detecting melodic segment boundaries. Finally, we wish to investigate whether there are further problems where our method could be beneficial. In general, pseudo-supervised optimization could improve features which are noisy either because of the way they are calculated, or because of noise in the data on which the features are based on.

Acknowledgments

The project Lrn2Cre8 acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859. We thank Marcus Pearce for sharing the Essen data used in [Pearce *et al.*, 2010a].

References

[Baddeley, 1966] A. D. Baddeley. Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, 18(4):362–365, 1966.

[Bregman, 1990] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[Brent, 1999] Michael R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. 34(1–3):71–105, 1999.

[Cambouropoulos, 2001] E. Cambouropoulos. The local boundary detection model (lbdm) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC'2001)*, Havana, Cuba, 2001.

[Frankland and Cohen, 2004] Bradley W Frankland and Annabel J Cohen. Parsing of Melody: Quantification and Testing of the Local Grouping Rules of Lerdahl and Jackendoff's A Generative Theory of Tonal Music. *Music Perception*, 21(4):499–543, 2004.

[Glicksohn and Cohen, 2011] Arit Glicksohn and Asher Cohen. The role of Gestalt grouping principles in visual statistical learning. *Attention Perception & Psychophysics*, 73:708–713, 2011.

[Gobet and Simon, 1998] F. Gobet and H. Simon. Expert chess memory: Revisiting the chunking hypothesis. *Memory*, 6:225–255, 1998.

[Gobet *et al.*, 2001] F. Gobet, P.C.R. Lane, S. Croker, P. C-H. Cheng, G. Jones, I. Oliver, and J.M. Pine. Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6):236–243, 2001.

[Goh *et al.*, 2010] H Goh, N Thome, and M Cord. Biasing restricted Boltzmann machines to manipulate latent selectivity and sparsity. *NIPS workshop on deep learning and unsupervised feature learning*, 2010.

[Grandvalet and Bengio, 2004] Yves Grandvalet and Yoshua Bengio. Semi-supervised Learning by Entropy Minimization. *Advances in Neural Information Processing Systems*, 17:529–536, 2004.

[Hinton *et al.*, 2006] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[Hinton *et al.*, 2014] Geoffrey E Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *NIPS 2014 Deep Learning and Representation Learning Workshop*, December 2014.

[Hinton, 2002] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, July 2002.

[Lattner *et al.*, 2015] Stefan Lattner, Maarten Grachten, and Carlos Eduardo Cancino Chacón. Probabilistic Segmentation of Musical Sequences using Restricted Boltzmann Machines. In *MCM 2015: Mathematics and Computation in Music: Proceedings of the 5th International Conference*, volume 9110, London, UK, 2015. Springer, Berlin.

[Lee, 2013] Dong-Hyun Lee. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML Workshop Challenges in Representation Learning WREPL*, pages 1–6, Atlanta, Georgia, 2013.

[Lerdahl and Jackendoff, 1983] Fred Lerdahl and Ray Jackendoff. *A generative theory of tonal music*. MIT press, 1983.

[McCollough and Vogel, 2007] Andrew McCollough and Edward Vogel. Visual chunking allows efficient allocation of memory capacity. *Journal of Vision*, 7(9):861, 2007.

[Meyer, 1956] L.B. Meyer. *Emotion and meaning in Music*. University of Chicago Press, Chicago, 1956.

[Narmour, 1990] E. Narmour. *The analysis and cognition of basic melodic structures : the Implication-Realization model*. University of Chicago Press, 1990.

[Nøklestad, 2009] Anders Nøklestad. *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection*. PhD thesis, University of Oslo, 2009.

[Pearce *et al.*, 2010a] Marcus Pearce, Daniel Müllensiefen, and Geraint A Wiggins. Melodic Grouping in Music Information Retrieval: New Methods and Applications. *Advances in Music Information Retrieval*, 274(Chapter 16):364–388, 2010.

[Pearce *et al.*, 2010b] Marcus T Pearce, Daniel Müllensiefen, and Geraint Wiggins. The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, 39(10):1367–1391, 2010.

[Pearce, 2005] M. T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, UK., 2005.

[Schaffrath, 1995] Helmut Schaffrath. The Essen Folksong Collection in Kern Format. In David Huron, editor, *Database containing , folksong transcriptions in the Kern format and a -page research guide computer database*. Menlo Park, CA, 1995.

[Taylor *et al.*, 2006] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. Modeling Human Motion Using Binary Latent Variables. *Advances in Neural Information Processing Systems*, pages 1345–1352, 2006.

[Temperley, 2001] D. Temperley. *The Cognition of Basic Musical Structures*. MIT Press, Cambridge, Mass., 2001.

[Tieleman and Hinton, 2009] T. Tieleman and G.E. Hinton. Using Fast Weights to Improve Persistent Contrastive Divergence. In *Proceedings of the 26th international conference on Machine learning*, pages 1033–1040. ACM New York, NY, USA, 2009.

[Tieleman, 2008] T. Tieleman. Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM New York, NY, USA, 2008.

[von Helmholtz, 2005] Hermann von Helmholtz. *Treatise on physiological optics*, volume 3. Courier Corporation, 2005.

[Wertheimer, 1938] Max Wertheimer. Laws of organization in perceptual forms. *A source book of Gestalt psychology*, pages 71–88, 1938.