DEVELOPING TONAL PERCEPTION THROUGH UNSUPERVISED LEARNING

Carlos Eduardo Cancino Chacón, Stefan Lattner, Maarten Grachten
Austrian Research Institute for Artificial Intelligence
{carlos.cancino,stefan.lattner,maarten.grachten}@ofai.at

ABSTRACT

The perception of tonal structure in music seems to be rooted both in low-level perceptual mechanisms and in enculturation, the latter accounting for cross-cultural differences in perceived tonal structure. Unsupervised machine learning methods are a powerful tool for studying how musical concepts may emerge from exposure to music. In this paper, we investigate to what degree tonal structure can be learned from musical data by unsupervised training of a Restricted Boltzmann Machine, a generative stochastic neural network. We show that even based on a limited set of musical data, the model learns several aspects of tonal structure. Firstly, the model learns an organization of musical material from different keys that conveys the topology of the circle of fifths (CoF). Although such a topology can be learned using principal component analysis (PCA) when using pitch-only representations, we found that using a pitch-duration representation impedes the extraction of the CoF topology much more for PCA than for the RBM. Furthermore, we replicate probe-tone experiments by Krumhansl and Shepard, measuring the organization of tones within a key in human perception. We find that the responses of the RBM share qualitative characteristics with those of both trained and untrained listeners.

1. INTRODUCTION

Modern approaches in music theory recognize that tonality can be broadly described as the organization of pitch classes into a hierarchical structure of tensions-relaxations around a tonal axis [10, 15, 16]. This conception of tonality is not limited to western tonal classical music, but can also be applied to modal music, popular music (e.g. jazz, rock) and non-western folk music [3]. This notion of tonality is not only a music theoretic construct: perceptual processing of musical stimuli in human listeners has been found to exhibit this type of organization as well [10]. Specific types of hierarchical organization of pitch classes are partly explained by acoustic attributes of pitch, especially the consonance between pairs of pitches [10], suggesting that low-level processing of acoustic stimuli may be relevant for the perception of tonal structure.

However, tonal structure is not only reflected in the physical attributes of pitch, it is also manifest in the statistical properties of music, such as the duration and frequency of occurrence of pitches [17], as illustrated in Figure 1. As Saffran et al. have shown [14], human listeners (including infants) are sensitive to such statistical regularities, and this leads to the view that tonal perception may be shaped by (long time) exposure to music exhibiting statistical regularities regarding frequency of occurrence of pitches, rhythmic emphasis, the position of occurrence within musical phrases, and possibly other aspects [9].

It is this process, the formation of tonal structure through exposure to musical stimuli, that we focus on in this paper. We choose a particularly straightforward approach, using a Restricted Boltzmann Machine (RBM) [6] to learn the probability distribution of melodic sequences, represented as n-grams of notes. In a first explorative experiment, we examine to what degree the feature space learned by the RBM is musically meaningful. Using the resemblance of the feature space to the circle of fifths as a quantitative criterion, we investigate the impact of the n-gram length, and compare pitch-only input representations to input representations that include both pitch and duration. In a second experiment, we use the RBM to simulate listener ratings in a probe tone test, and compare the results to ratings from human listeners of different skill levels.

The structure of the paper is as follows: In Section 2, we discuss prior work on the induction of tonal structure using computational models. Section 3 relates the different aspects of the unsupervised learning task to various perceptual mechanisms that are assumed to be at play in the perception of tonal structure. Section 4 briefly describes the RBM model, the data used for training the model, and representation of the data. The experiments on tonal organization and the organization of pitches are described in Sections 5 and 6, respectively. Conclusions and future directions are presented in Section 7.

2. RELATED WORK

The idea of studying the perception of tonal structure by using computational models to simulate the enculturation process is not new. For example, Tillmann et al. [18] use a hierarchical self-organizing map (SOM) [8] to learn representations of tonal structure from pitch-class representations of chord sequences. They find that their model is able

to develop an organization comparable to that of empirical data gathered from various studies on human perception of tonality. Leman [12] presents an alternative approach to modeling the perception of tonality. He employs a psychoacoustic model in combination with a SOM to learn tonal representations starting from acoustic data. Furthermore, Toiviainen & Krumhansl examined the perception of musical scales by projecting human ratings to the feature space of a SOM, which was trained on scale profiles of Krumhansl [19].

A commonality among the mentioned works is the choice of the self-organizing map as a model for accommodating the learning process. The reason for this preference may be that both the spatial mapping of the data, and the competitive learning algorithm employed by the SOM, are biologically plausible characteristics of the human sensory cortex [7]. The RBM model used in the work presented here, is not explicitly presented as (nor was it designed to be) a biologically plausible model of learning in the brain. Nevertheless RBMs and deep belief nets based on RBMs, in combination with sparseness constraints on the activation of hidden units, are able to learn features from visual data that strongly resemble receptive fields of neurons in the visual cortex [11]. As such, RBMs prove to be a valid computational modeling approach for learning biologically plausible representations from musical data.

A fundamental difference between SOMs and RBMs is that in the former, the hidden units represent points in an explicitly defined low-dimensional feature space. In RBMs, the feature space is defined by the set of all possible combinations of hidden unit activations, such that each hidden unit represents a dimension of the feature space. This allows for representations of data instances as a (non-linear) combination of features. The topology of this high-dimensional feature space can be visualized in a 2-D space using PCA.

3. PERCEPTUAL MECHANISMS

As argued by Smith and Schmuckler [17], perceptual processes like *discrimination*, *differentiation* and *organization* play an important role in the perception of musical tonality. In this Section, we will briefly describe these processes, and show how they can be related to formal aspects of the computational modeling methods, such as the choice of input data representation, and the topology of the feature space being learned.

Perceptual discrimination refers to the sensitivity of a system to differences along some perceivable stimulus dimension. In computational learning models, this relates to the form of input data representation. In general, the type of relevant input features depends heavily on the respective learning task [1]. Musical data comprises much context-dependent information that can not be trivially inferred from low-level representations. To decide on an appropriate representation is thus not always an easy task. For instance, pitch content can be represented in several ways, such as frequency spectra, MIDI note numbers, or pitch classes. In our current experiments, we use MIDI note numbers as well as pitch class representations. Duration is encoded separately from pitch. An advantage of this over combined pitch-duration representations (e.g. piano-roll notation) is that the n-gram size is specified in the number of notes, rather than an absolute time interval. This allows for comparing pitch-only to pitch-duration representations. The input data will be referred to as Input Space (IS), and will be described in more detail in Section 4.3.

Differentiation is a higher order ability that refers to the segregation of the perceived stimuli into elements on the basis of its discriminable differences [17]. In an unsupervised model we can identify this ability as the capacity of the system to segregate the data in the IS into clusters in the Feature Space (FS). In the context of tonality, an example of differentiation would be the capacity of an unsupervised model to cluster the data in the FS in such a way that each cluster represents a musical key. A measure of quality of this clustering would then be the variance of each cluster, as smaller variances imply a better differentiation of the data with respect to each class.

Organization builds on the concept of differentiation, as it establishes relations between the differentiated elements, as well as the nature of the relations themselves. In an unsupervised model, this can be understood as the topology of the FS. In this way, geometric features such as the distance between clusters, as well as the relative position between them can express similarity.

Bharucha [10, cited by Krumhansl] recognizes two types of hierarchies regarding musical tonality. *Event hierarchies* refer to the functional significance of single note events in a specific musical context, while *tonal hierarchies* account for the abstract musical structure in a particular culture or genre, e.g. the functional significance of all elements of a pitch class relative to all other pitch classes.

In our case, we compare the organization of the data in the FS to the circle of fifths, a well known music theoretical construct that explains the relations and the neighborhood of keys [15]. As a measure of quality we use the Procrustes Distance (PD) [4] of the centroids of the clustered data in the feature space with respect to the CoF.

4. METHODS

4.1 Restricted Boltzmann Machine

A Restricted Boltzmann Machine is a stochastic Neural Network (NN) with two layers, a visible layer with units $\mathbf{v} \in \{0,1\}^r$ and a hidden layer with units $\mathbf{h} \in \{0,1\}^q$ [6]. 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. Given a visible vector $\mathbf{v}$, the free energy of the model can be calculated as:

$$\mathcal{F}(\mathbf{v}) = -\mathbf{a}^\intercal \mathbf{v} - \sum_i \log\left(1 + e^{(b_i + \mathbf{W}_i \mathbf{v})}\right), \quad (1)$$

where $\mathbf{a} \in \mathbb{R}^r$ and $\mathbf{b} \in \mathbb{R}^q$ are bias vectors, and $\mathbf{W}_i$ is the i-th row of the weight matrix.

Given $\mathbf{v}$, a sample of $\mathbf{h}$ can be obtained from its conditional activation probability, given by:

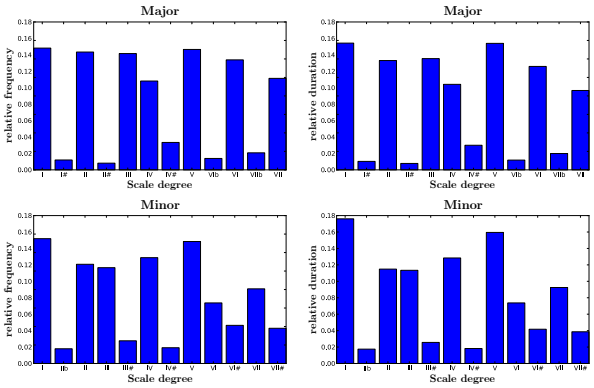

Figure 1. Occurrence and duration distributions of the fugues from Bach's Well Tempered Clavier.

$$p(\mathbf{h} = \mathbf{1} \mid \mathbf{v}) = \sigma\left(\mathbf{b} + (\mathbf{v}^\intercal \mathbf{W})^\intercal\right), \qquad (2)$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic sigmoid function.

In experiment 1, we consider the conditional activation probability of vector $\mathbf{h}$ as the result of the projection of $\mathbf{v}$ into the FS. In the second experiment, we calculate the energy using Eq. (1).

4.1.1 Training

We train the model with 200 hidden units for 1000 epochs with Contrastive Divergence (CD) [6], using 3 Gibbs sampling steps and a mini-batch size of 500 for the weight updates. The learning rate is set to 0.01 and the momentum to 0.3. These parameters were empirically selected according to the rules of thumb suggested by Hinton in [5]. In addition, we use the well-known L2 weight-decay regularization which penalizes large weight coefficients.

Based on properties of neural coding, sparsity and selectivity can be used as constraints for the optimization of the training algorithm [2]. Sparsity encourages competition between hidden units, and selectivity prevents overdominance by any individual unit. These constraints are used in our training, with a linear falloff of its influence over the first 200 epochs from 50% to 30%.

4.2 Training Corpus

J. S. Bach's Well Tempered Clavier (WTC), composed between 1722 and 1742, is widely recognized as one of the most influential works in music history [15]. It is also one of the most important works that systematically spans the whole range of major and minor keys, and is therefore well-suited for experiments on tonality. In this paper, we use MIDI versions of the 48 fugues of the WTC as corpus, encoded by David Huron and taken from the KernScores website (http://kern.ccarh.org). Each fugue is decomposed into its voices (two to five), and we consider each voice as a single monophonic melody in its respective key. In Figure 1, the distributions of the occurrence and duration of the notes of the WTC are shown. These distributions are similar to the key profiles by Krumhansl & Kessler [19].

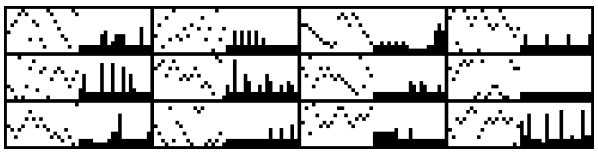

Figure 2. Twelve random *pitch-duration* training instances of the WTC corpus as 20-grams before linearization. Notes are ordered horizontally, the vertical dimension accounts for pitch and duration values, respectively. The left part of each instance shows the one-out-of-m pitch representation of 20 consecutive notes, the right part shows the corresponding duration representation.

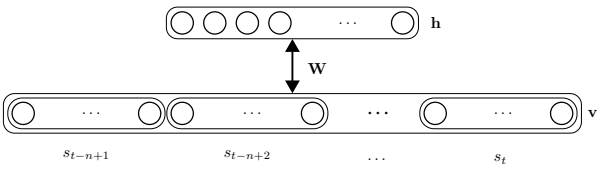

Figure 3. The RBM architecture used. An input vector $\mathbf{v}$ is constituted by a linearized n-gram, where s_j is a binary representation of note j.

4.3 Input 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. Depending on the experiment, we either use only pitch information, or we use both the pitch and duration of the notes. In the first case, an n-gram is a concatenation of n bit vectors of size m, where the i-th bit vector is a one-out-of-m representation of the pitch of note i.

In the second case, n additional vectors are added to the n-gram, where the i-th vector now represents the duration of the i-th note (see the right half of the instances shown in Figure 2). Such a duration vector is constructed by quantizing all durations of a melody into 12 bins and by relating each of those to one of 12 units. A duration that falls into bin k is represented by activating units 1 to k. After linearization, the resulting n-gram constitutes the visible vector $\mathbf{v}$, as illustrated in Figure 3.

5. TONAL ORGANIZATION

In this experiment, we examine the ability of an RBM to learn tonal relationships between n-grams. To that end, we project the FS learned by the RBM into a two-dimensional space using Randomized Principal Component Analysis (rPCA) [13]. As the CoF is the underlying music theoretical construct for the relationships between keys, we are interested to what degree we can approximate the CoF topology. As a baseline, we compare this projection to a direct projection of the IS, again using rPCA.

5.1 Training

We encode the WTC corpus as described in 4.3. As keys are characterized by distributions of pitch classes, the pitch range is set to $m = 12$. In order to examine the organization ability of the RBM under different settings, we use

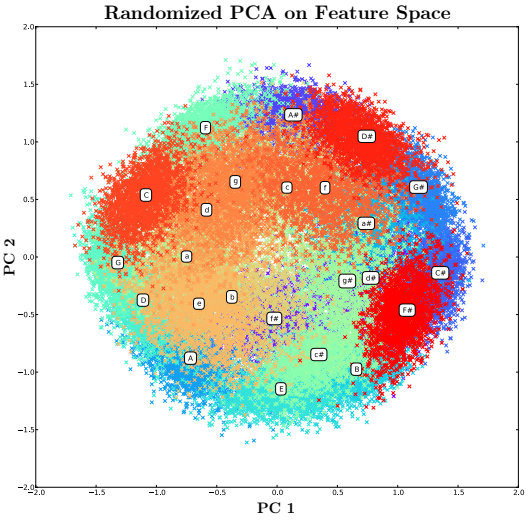

Figure 4. 2-D visualization of n-grams in the FS using rPCA. N-grams belonging to a key have the same color, each centroid is marked with the corresponding cluster's key label. (Best viewed in color)

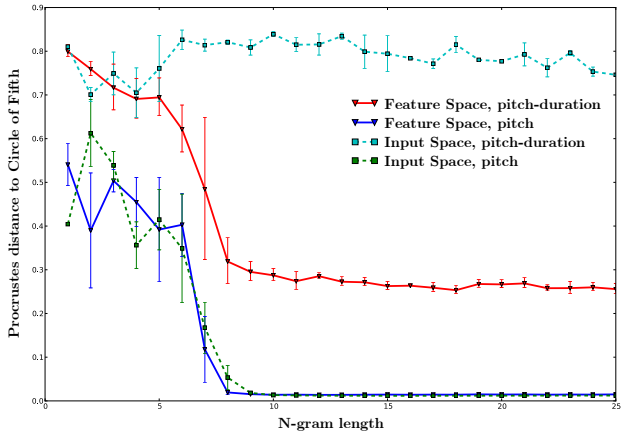

Figure 5. The average Procrustes Distances from major key centroids to the major CoF of 5 runs for different n-gram lengths after rPCA on the IS and on the FS. *pitch* and *pitch-duration* representations are used as input.

n-grams of various lengths, and also compare *pitch* and *pitch-duration* representations.

5.2 Evaluation

We use rPCA to project all n-grams in both the IS and the FS into a two-dimensional space. In this space, for each key we determine the mean of all n-grams created from pieces in that key. The organization of those centroids is then compared to the organization of keys in the CoF by computing the PD of both shapes, separately for major and minor keys. To make different expansions of data points in space comparable, the PD is finally divided by the perimeter of the target CoF.

5.3 Results and Discussion

Figure 4 shows the organization of n-grams in the FS. Cluster centers are organized similarly to how keys are organized in the CoF, which is consistent with the representations of the probe tone ratings obtained by Krumhansl and Kessler [9, 10]. Note that relative minors tend to be shifted counterclockwise with respect to their major counterparts. This occurs in Krumhansl's results as well [10, pp. 43], and can be explained by two factors, namely the alteration of the sixth degree in the melodic minor scale, which is identical to the seventh degree of the dominant of the relative major counterpart (e.g. the melodic Am scale shares the F# with the G major scale, the dominant of C major), and due to the tonal modulations concerning the form of the piece (e.g. fugues in minor keys tend to have certain passages in the relative major, while fugues in major keys tend to have passages in both the relative minor and the dominant).

Figure 5 shows, that the Procrustes Distance to the CoF tends to stabilize at a minimum with an n-gram length of about nine. This can be explained by the fact that n-grams of that length contain enough information to obtain the respective distribution of a key well enough. Adding duration information clearly impedes the organization of clusters in a CoF topology. As the occurrence of notes in the WTC is strongly correlated to their absolute duration (see Figure 1), and rhythmic information is not directly linked to the CoF organization, this is not unexpected. Interestingly, for larger n-gram sizes the FS of the RBM is not disrupted as much by the inclusion of distractive information as the rPCA on the IS.

6. ORGANIZATION OF PITCHES

A probe-tone test, proposed by Krumhansl et al. [9, 10], consists of a set of *musical stimuli* (such as scales, chord cadences, or musical pieces) that unambiguously instantiate a specific key, and a set of *probe tones*, typically the set of 12 pitch classes. Listeners are then required to rate on a numerical scale, from 1 ("very bad") to 7 ("very good"), how well the probe tones fit the musical stimulus. In order to explore the hierarchical event organization of pitches induced by the RBM, we compare our model with a particular probe tone test conducted by Krumhansl and Shepard [10, cited by Krumhansl]. In this specific experiment, the musical stimulus consisted of an incomplete C major scale (in both ascending and descending contexts), and listeners were asked to give a numerical rating of the degree to which each probe tone fits the scale. The stimuli of this particular setup are illustrated in part Figure 6 a), while the probe tones are shown in Figure 6 c). The participants of the experiment were divided in three groups according to their number of years of formal musical training.

6.1 Training

As we are only interested in the ability of the model to learn tonal hierarchies in major and minor mode, we transpose all melodies to C major and C minor, respectively. In order to remain consistent with the aforementioned experiment of Krumhansl & Shepard, rather than using pitch-classes, we allow the training data to be in a range of three octaves, ranging from MIDI pitch numbers 48 to 74 (such that both the stimuli and the probe tones can be represented

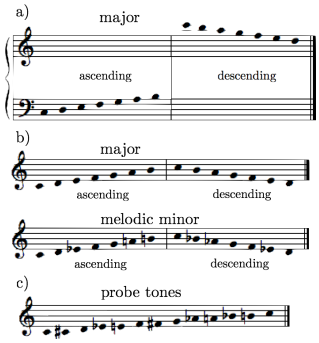

Figure 6. Stimuli/probe tones used in the probe-tone test.

without wrapping). Most of the n-grams of the transposed WTC data fall in that range, or can be transposed octave-wise to fall in that range. N-grams for which this is not possible are ignored.

Since the fugues from the WTC contain certain tonal modulations, in order to train the RBM with prototypical examples of major and minor scales, all n-grams are classified using the Krumhansl & Kessler key-finding algorithm [10, cited by Krumhansl] and those whose annotated key is not the same as that identified by the classifier (ca. 53% of the corpus) are removed. The training is executed as described in 4.1.1.

6.2 Evaluation

Two different probe tone tests are conducted. The first test aims to reproduce the setup by Krumhansl and Shepard, and thus, the stimuli consist of the major ascending (starting from C3) and descending scales (starting from C5) shown in Figure 6 a). For the second experiment, the stimuli consist of ascending and descending major and melodic minor scales, but this time both are generated in the middle C octave, as shown in Figure 6 b). For both tests, the set of probe tones consist of all notes of the chromatic scale (starting from C4) as shown in Figure 6 c). We construct n-grams of length 8, consisting of the 7 notes of the target stimulus and a probe tone as the last note. This results in visible vectors $\mathbf{v}_{pt}$ of length 36×8. The free energy corresponding to each combination of stimulus and probe tone is calculated using Eq. (1). In order to compare our results to those of human listeners, these energies are scaled using an affine transformation as follows:

$$\text{Judgment}(\mathbf{v}_{pt}) = \alpha \mathcal{F}(\mathbf{v}_{pt}) - \beta, \qquad (3)$$

where the constants α, β are selected such that the mean and the variance of the scaled energy are equal to those of the judgments reported in [10].

6.3 Results and Discussion

Figure 7 shows the results of the probe tone test, and in Table 1 the correlations of the RBM judgments with respect to those of expert and untrained listeners are presented. These results suggest that the model can learn some event hierarchy structures, such as the prevalence of diatonic over chromatic notes, similar to the judgment of trained listeners. In addition, the model develops a sense for melodic direction, preferring probe tones close to the final notes of the stimulus, which is consistent with the ratings of untrained listeners. Stimulated in the middle octave, the model is able to distinguish major and minor modes, especially the major and minor thirds reflect the characteristics of the respective diatonic triads. The model responses do not show explicit octave equivalence, since C and C' are not equally emphasized. Still it is interesting to note that a stimulus in the lower octave has implications on the pitch expectations in the middle octave, and that these implications are in correspondence with the tonal hierarchy of the key implied by the stimulus.

Group	r	p-value
Expert ascending	0.7213	0.0054
Untrained ascending	0.7942	0.0012
Expert descending	0.7985	0.0011
Untrained descending	0.8344	0.0004

Table 1. Pearson correlations and p-values for the judgments of the probe tone tests.

7. CONCLUSION

In this paper we show that tonal structure can be learned from musical data with an RBM using unsupervised training with a limited set of monophonic melodies. The model is able to reproduce the topology of the CoF using *pitch* n-gram representations of the input data. We found that for successful inference of the CoF, a minimal n-gram length of nine notes is needed, and that longer n-grams do not lead to better representations. Furthermore, although duration information profoundly disturbs the learning of tonal structure through the baseline rPCA method, the RBM model is less affected by distracting duration information.

By way of a probe tone test, we explored the organization of pitches in the context of major and minor modes. Our results show the model was able to learn several aspects of tonal structure, in particular the hierarchical prevalence of diatonic over chromatic tones. Comparing results with Krumhansl's probe tone experiments on human subjects with different levels of musical training do not yield a conclusive classification of the model: the model displays aspects of both untrained and trained subjects.

An important feature of tonal perception in trained subjects is octave equivalence. This feature was not well-reproduced by the model. It is possible that a pre-condition for octave-equivalence is the harmonic overlap of octaves. In our current setup, the overtone structure of tones is not represented. To test this hypothesis, we intend to investigate whether using harmonic tone representations leads to stronger octave-equivalence in the the model.

Furthermore we wish to investigate which factors induce more expert-like perception of tonal structure. Possible factors include the size of the training data, and the depth of the model (in terms of hidden layers).

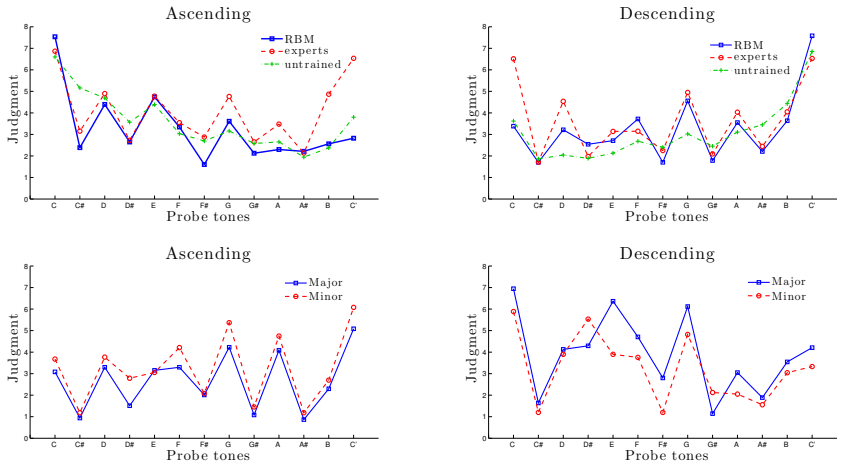

Figure 7. (Top) Comparison of the judgments for the probe tones between the RBM and human listeners for both ascending (left) and descending (right) major stimulus in the lower and upper octave, respectively. (Bottom) Comparison of the judgments for the probe tones of the RBM for both major and melodic minor stimulus in the middle octave. In all cases, responses are measured in the middle octave.

8. ACKNOWLEDGMENTS

This work is supported by the European Union Seventh Framework Programme, through the Lrn2Cre8 project (FET grant agreement no. 610859). We thank Geraint Wiggins, Kat R. Agres and Jamie Forth for valuable suggestions and commentaries on this work.

9. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2009.

[2] 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*, pages 1–8, 2010.

[3] E. Gómez. *Tonal description of music audio signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2006.

[4] C. Goodall. Procrustes Methods in the Statistical Analysis of Shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991.

[5] G. E. Hinton. A practical guide to training restricted Boltzmann machines. Tech. Report UTML TR 2010-003, Department of Computer Science, University of Toronto, 2010.

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

[7] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[8] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybernetics*, 43:59–69, 1982.

[9] C. L. Krumhansl and L. L.Cuddy. A theory of tonal hierarchies in music. In *Handbook of Auditory Research*. Springer, New York, 2010.

[10] C. L. Krumhansl. *Cognitive foundations of musical pitch*. Cognitive foundations of musical pitch. Oxford University Press, New York, 1990.

[11] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area V2. In *Advances in Neural Information Processing Systems 20*, pages 873–880. 2008.

[12] M. Leman. A model of retroactive tone center perception. *Music Perception*, 12(4):439–471, 1995.

[13] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *arXiv.org*, page 2274, 2008.

[14] J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27 – 52, 1999.

[15] F. Salzer. *Structural hearing; tonal coherence in music*. New York, Dover Publications, 1962.

[16] H. Schenker. *Harmony*. University of Chicago Press, 1980.

[17] N. A. Smith and M. A. Schmuckler. The Perception of Tonal Structure Through the Differentiation and Organization of Pitches. *Journal of Exp. Psych.: Human Perception and Performance*, 30(2):268–286, 2004.

[18] B. Tillmann, J. J. Bharucha, and E. Bigand. Implicit learning of tonality: a self-organizing approach. *Psychological review*, 107(4):885–913, 2000.

[19] P. Toiviainen and C. L. Krumhansl. Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, 32(6):741–766, 2003.