LEARNING A ROBUST TONNETZ-SPACE TRANSFORM FOR AUTOMATIC CHORD RECOGNITION

Eric J. Humphrey, Taemin Cho, and Juan P. Bello

Music and Audio Research Laboratory (MARL)
New York University, New York USA
{ejh333, tmc323, jpbello}@nyu.edu

ABSTRACT

Temporal pitch class profiles – commonly referred to as a chromagrams – are the de facto standard signal representation for content-based methods of musical harmonic analysis, despite exhibiting a set of practical difficulties. Here, we present a novel, data-driven approach to learning a robust function that projects audio data into Tonnetz-space, a geometric representation of equal-tempered pitch intervals grounded in music theory. We apply this representation to automatic chord recognition and show that our approach out-performs the classification accuracy of previous chroma representations, while providing a mid-level feature space that circumvents challenges inherent to chroma.

Index Terms— Chord Recognition, Deep Learning, Convolutional Neural Networks, Tonnetz

1. INTRODUCTION

As evidenced by its established research history in the music informatics community, the harmonic analysis of digital musical content is an active research area with myriad applications. Common tasks include automatic chord transcription, song segmentation and, more recently, structural analysis, to name a few popular topics from the field.

For nearly as long as there has been interest in this class of applications, pitch class profiles, or chromagrams, stand alone as the predominant input feature for harmonic analysis systems. Though much progress has been realized based on the frame-wise analysis of chroma, the representation presents a set of practical challenges. Defined as the projection of the pitch helix to fundamental pitch class by discarding height (octave) information, as illustrated in Figure 1, chroma is naturally sensitive to variations in the signal-level attributes of timbre and loudness. Additionally, moving voices (e.g. bass lines and melodies) may manifest awkwardly in these representations, and significant effort has been invested in reducing the effects of such nuances [1] [2].

Furthermore, a chromagram representation exhibits the properties of a discrete distribution, which proves cumbersome when used as a mid-level feature space. The spatial configuration of the pitch classes within the distribution is inconsequential, and therefore the distance between two chroma vectors only captures the activation differential. The circular nature of pitch is also not natively encoded in a chroma distribution, and the vector must be explicitly rotated to identify key modulations or other forms of intervalic motion.

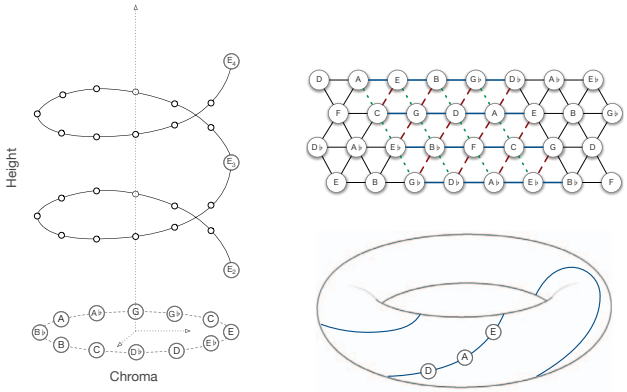

Fig. 1. Theoretical Models of Pitch – the pitch helix (left) and the Tonnetz (right). When equal-temperament tuning is imposed, pitch class is equivalent regardless of octave, and both representations wrap to circular models. As shown in the planar Tonnetz (top right), the relationship between pitch is defined by fifths (red solid), minor thirds (dotted green) and major thirds (dashed blue).

Conversely, the Tonnetz is a harmonic network representation of pitch intervals where, by imposing an equal temperament tuning system (octave equivalence), the surface of the model wraps onto itself as a six-dimensional hypertorus [3], illustrated in Figure 1. The primary advantage gained by adopting a Tonnetz representation of tonality is that pitches and chords live in a geometric space, and therefore distances between points are musically meaningful. Additionally, the Tonnetz explicitly encodes interval information and lends itself well to chord-based applications. Due to the latent geometry of the model, intervalic analysis can be achieved by differentiating a given trajectory – like removing a DC-bias

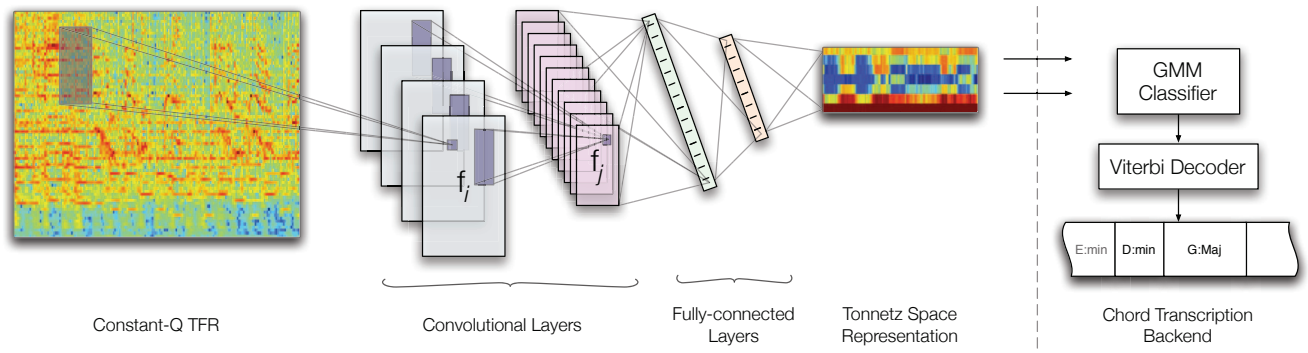

Fig. 2. Full System.

– a potentially desirable attribute in the context of computing harmonic similarity.

To address the two challenges stated previously, we propose a data-driven method utilizing deep learning techniques to produce a Tonnetz-space transformation that automatically extracts harmonic, signal-level features while being insensitive to fluctuations in timbre or loudness. Though such techniques have recently been applied to instrument retrieval [4] and genre classification [5], we are unaware of any instances in the area of harmonic analysis or automatic chord recognition. The remainder of the paper is organized as follows: Section 2 presents the proposed system architecture and both input and output representations, Section 3 addresses our evaluation methodology, we present and discuss our results in Section 4 and provide conclusions and directions for future work in Section 5.

2. PROPOSED SYSTEM

We present an approach to learning a relatively short-time scale transformation from high-dimensional time-frequency audio representations to a low-dimensional output space grounded in music theory. As shown in Figure 2, the system that projects data into the Tonnetz space is comprised of an appropriately chosen input time-frequency representation, a convolutional neural network architecture, and the output representation corresponding to this model.

2.1. Time-Frequency Representation

Given that linear pitch shifts are logarithmic in frequency, the input time-frequency representation must be warped in frequency to allow the network to learn translation-invariant features, i.e. intervals. This is achieved with the Constant-Q transform, which serves as a base-2 logarithmic downsampling operation in frequency and can be efficiently implemented as a complex matrix multiply with the short-time Fourier transform (STFT).

Audio signals are downsampled to 11025Hz and transformed by the STFT with frame and analysis hop sizes of 8192 and 1024, respectively, resulting in a frame rate of 10.77Hz. The modified Constant-Q kernel is generated with 78 filters spanning 41–3520Hz at 12 bins per octave. In previous work [4], we observed that narrowing the filters by raising the conventional Hamming windows to the 8^{th} power and normalizing the magnitude responses reduces the smearing of pitch information across adjacent coefficients, enhancing tonal information.

2.2. Architecture

As timbre and loudness vary a great deal in reality, it is exceedingly difficult to arrive at a Tonnetz-space representation by a naïve reduction of time-frequency information. Therefore, we aim to produce a robust function that, by automatically extracting features and encoding this information in the transform, is resilient to variations in both timbre and loudness. Importantly, since traditional, fully-connected neural networks are incapable of capturing spatial correlations in data, we use convolutional neural networks (CNN) to avoid this issue. Pioneered by LeCun et al in [6], CNNs offer sensitivity to spatially correlated data and invariance to feature translation.

CNNs are an extension of classical discriminant functions where weights are shared across an input vector as spatially shifted convolutions, and are often followed by a downsampling operation, the addition of a bias term and application of a non-linear activation function; for clarity, we refer again to Figure 2. Collectively, each layer of a CNN, shown horizontally, maintains several sets of weights, or kernels, that are convolved with its inputs, producing a set of output vectors called feature maps, marked as f_i and f_j. Successive convolutional layers are capable of learning hierarchical, shift-invariant features, and are powerful enough to be applied directly to raw input representations.

We define the input to the CNN to be a 'tile' of 3 frames (roughly 300ms) by 78 constant-Q coefficients and construct

a network of 2 convolutional layers and 3 fully-connected layers. The first layer has 30 kernels with a (2×7) shape; similarly, the second layer has 36 kernels with a (2×13) shape. The fully connected layers consist of 252, 42, and 7 units. In total, the architecture has 1,726 parameters (weights and biases) and two model hyper-parameters (the learning rate, $\eta = 0.033$, and weight decay $\lambda = 0.01$). The machine is implemented using the Theano package [7] developed by the LISA Machine Learning Laboratory at the University of Montreal. Notably, downsampling fields along frequency dimensions in this architecture are explicitly avoided, as such a process will introduce tonal ambiguity.

2.3. Tonnetz Space

Referring to a more comprehensive review in [3], we define a 7-dimensional output space, corresponding to the 6-D Tonnetz space and a 1-dimensional null-chord (no chord) regressor. The origin of the Tonnetz-space is a poor choice for the null-chord target, as the machine would attempt to encode amplitude information along the radius of the space, distorting tonal information. Centroids of the 12 major and minor triads are calculated as training templates, and the null-dimension is defined as 1 and -1 for positive and negative chord instances, respectively.

3. METHODOLOGY

To characterize the usefulness of the learned Tonnetz projection, we describe the data used for training and evaluation, the process of training the parameterized function and the defined loss function, as well as outline the chord detection task.

3.1. Data

Acknowledging the common heuristic that neural networks typically require large amounts of data to train adequately, we conduct our work on a set of 493 chord-annotated polyphonic sound recordings, consisting of 179 songs from Christopher Harte's Beatles dataset, 20 songs from Matthais Mauch's Queen dataset, 100 songs from the RWC Pop dataset and 194 songs from the US Pop dataset, as described in [8]. All chord labels are quantized to the 12 major and 12 minor triads, in addition to the null-chord label. For comparison purposes, we perform 13-fold cross validation across each album in the Beatles set for testing. For each of the thirteen holdout scenarios, the training data is split into 10 folds and training repeated five times.

3.2. Training

Different models were trained for all combinations of training folds and holdout datasets, proceeding by presenting batches of 125 randomly selected tiles from the training set and performing mini-batch stochastic gradient descent over the weights $\mathcal{W}$ with hyperparameters η, the update rate, and λ, the weight decay of the L_1 regularizer, given in Equation (1).

$$\mathcal{W} \leftarrow \mathcal{W} - \eta * (\frac{\partial \mathcal{L}}{\partial \mathcal{W}} + \lambda \sum_k \|W_k\|_1) \tag{1}$$

All training sessions were run for a minimum of 20,000 iterations, and it was observed early in the training of these models that introducing the regularizer over the kernel weights accelerated convergence. Without imposing sparsity constraints, it was not uncommon for the training process to idle for many iterations. The values of the mini-batch size and update rate were tuned empirically until the training process converged consistently.

To simultaneously learn the 6-D Tonnetz projection Z_d and the 1-D null-chord regressor Z_{null}, we define a contrastive loss functional as (2). This function alternates between the Tonnetz loss (3) and the null-chord loss (4), based on the sign of of the target T_{Null}, where Equation 5 is a differentiable approximation of the step function, tuned by the parameter Q.

$$\mathcal{L} = L_{Ton} + L_{Null} \tag{2}$$

$$L_{Ton} = \frac{1}{2} Y * (\sum_{d=1}^{6} (Z_d - T_d)^2 + (Z_{Null} - 1)^2) \tag{3}$$

$$L_{Null} = \frac{1}{2}(1 - Y) * (Z_{Null} + 1)^2 \tag{4}$$

$$Y = sigmoid(Q * T_{Null}) \tag{5}$$

3.3. Experiments

Applying the learned Tonnetz transform to automatic chord recognition, we use a multivariate Gaussian Mixture Model (GMM) classifier of six Gaussians with diagonal covariance matrices and perform smoothing via the Viterbi decoder [2]. While we theoretically know the target locations of chords in the Tonnetz space beforehand, having defined these ourselves, we use a GMM classifier to compensate for any centroid drift that may arise as a result of chord label quantization. Even in the scenario where all chords deemed equivalent do in fact project to the same proximity at the output, a multivariate Gaussian model reduces to an over-complete k-means classifier. A transition penalty is applied to the Viterbi algorithm to tune the self-transition probability relative to all other chord transistions. Accuracy is defined as the percentage of total correct chord label duration over the duration of the dataset.

4. RESULTS AND DISCUSSION

As shown in Table 1, the learned Tonnetz transformation performs competitively with the state of the art in chroma-enhancement methods, outperforming the best system pre-

Table 1. Chord Recognition Statistics

Method	Accuracy
Tonnetz	**78.41**
Optimal Chroma Filtering [2]	75.7

sented in [2]. In addition to the high cumulative classification accuracy of 78.41%, compared to 75.70%, the average standard deviation of accuracy across folds is 0.19, indicating that the network consistently generalizes well. Chord confusions for both the system discussed here and the chroma features presented in [2] are shown in Figure 3. Overall, the learned Tonnetz representation confuses fewer chord types, except for the parallel minor in the Major chord case.

While performance varied across each holdout set as a function of the chord labels, harmonic content and production techniques, there was one notable outlier ("Lovely Rita"), which seldom eclipsed 6% accuracy. Upon closer inspection, it was apparent that there is an intonation discrepancy, as the detected chords are consistently flat by a semitone. Being that the learned transform does not attempt to correct for tuning inaccuracies, this behavior is identified as an area to explore in future work.

5. CONCLUSIONS

In this paper, we present an approach to using deep learning techniques to yield a function that projects high-dimensional data into a low-dimensional metric tonal space. This transformation produces an output representation that out-performs state of the art chroma filtering methods for chord recognition tasks, while the Tonnetz representation itself provides a set of desirable properties. Future work includes characterizing the system across a variety of parameters, such as kernel size in both time and frequency, network capacity or sparsity constraints on the network's weights, as well as tuning modifications to the input TFR. Finally, this approach does little to directly incorporate musical time or longer time-scale information. Chroma enhancement methods like recurrence-plot filtering [8] increase chord detection accuracy to over 80%, and informal experimentation with the application to a Tonnetz representation encourages improvement in accuracy as well.

6. REFERENCES

[1] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 3, pp. 649–662, 2010.

[2] T. Cho, R. J. Weiss, and J. P. Bello, "Exploring Common Variations in State of the Art Chord Recognition Systems," in *Proc. Sound and Music Computing Conference (SMC)*, Barcelona, Spain, July 2010, pp. 1–8.

[3] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, New York, NY, USA, 2006, AMCMM '06, pp. 21–26, ACM.

[4] E. J. Humphrey, A. Glennon, and J. P. Bello, " Non-Linear Semantic Embedding for Organizing Large Musical Instrument Sample Libraries," in *To appear in Proc. International Conference for Machine Learning and Applications (ICMLA)*, Honolulu, HI USA, Dec. 2011.

[5] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun, " Unsupervised Learning of Sparse Features for Scalable Audio Classification," in *To appear in Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL USA, Oct. 2011.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient based learning applied to document recognition.," *Proceedings of IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010, Oral.

[8] T. M. Cho and J. P. Bello, "A Feature Smoothing Method For Chord Recognition Using Recurrence Plots," in *To appear in Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL USA, Oct. 2011.

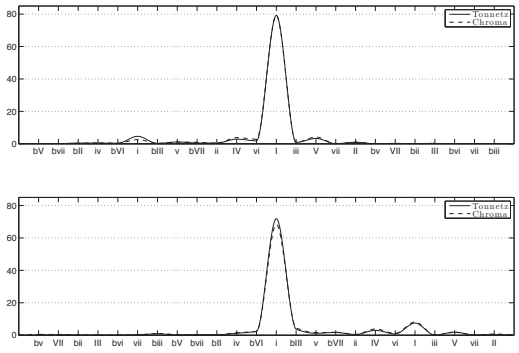

Fig. 3. Chord confusions for all Major (top) and Minor (bottom) chords, shifted to relative interval.