IMPROVED ONSET DETECTION FOR TRADITIONAL IRISH FLUTE RECORDINGS USING CONVOLUTIONAL NEURAL NETWORKS

Islah Ali-MacLachlan, Carl Southall, Maciej Tomczak, Jason Hockman
DMT Lab, Birmingham City University
islah.ali-maclachlan, carl.southall, maciej.tomczak, jason.hockman
@bcu.ac.uk

ABSTRACT

The usage of ornaments is key attribute that defines the style of a flute performances within the genre of Irish Traditional Music (ITM). Automated analysis of ornaments in ITM would allow for the musicological investigation of a player's style and would be a useful feature in the analysis of trends within large corpora of ITM music. As ornament onsets are short and subtle variations within an analysed signal, they are substantially more difficult to detect than longer notes. This paper addresses the topic of onset detection for notes, ornaments and breaths in ITM. We propose a new onset detection method based on a convolutional neural network (CNN) trained solely on flute recordings of ITM. The presented method is evaluated alongside a state-of-the-art generalised onset detection method using a corpus of 79 full-length solo flute recordings. The results demonstrate that the proposed system outperforms the generalised system over a range of musical patterns idiomatic of the genre.

1. INTRODUCTION

Figure 1: Player with Rudall and Rose eight-key simple system flute manufactured from cocus wood.

Irish Traditional Music (ITM) is a form of Folk music that developed alongside social dancing and has been an integral part of Irish culture for hundreds of years (Boullier, 1998). ITM consists of various subgenres and is played with a wide variety of traditional instrumentation, including melody instruments such as fiddles, bagpipes, tin whistles, accordions and flutes. Figure 1 presents an ITM performer with a wooden simple system flute.

Determining the stylistic differences between players is an important first step towards understanding how the music and culture associated with ITM has developed. Within traditional music, mastery is determined by technical and artistic ability demonstrated through individuality and variation in performances. Individual playing style is comprised of several features, including variations in melody, rhythmic phrasing, articulation, and ornamentation (McCullough, 1977; Hast & Scott, 2004; Keegan, 2010; Köküer et al., 2014).

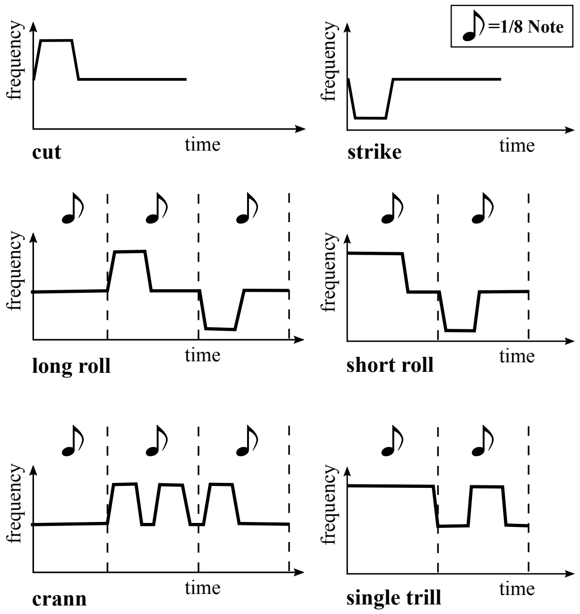


Figure 2: Frequency over time of *cut* and *strike* articulations showing change of pitch. *Long* and *short rolls*, *canns* and *single trills* are also shown with pitch deviations. Eighth-note lengths are shown for reference.

Automated identification of a player's style would be useful in the musicological investigation of various trends within the ITM timeline. A first step towards automated style identification is the detection of onsets related to notes and ornaments. This study continues the work of Ali-MacLachlan et al. (2016) by evaluating notes and single-note ornaments known as *cuts* and *strikes*. We also investigate breaths and the cut and strike elements of multi-note ornaments known as *short roll*, *long roll*, *crann* and *single trill* as described in Larsen (2003). Figure 2 depicts single-note and multi-note ornaments over time.

Onset detection algorithms are used to identify the start of musically relevant events. Ornament onset detection for Irish traditional flute recordings is a difficult task due to their subtle nature; ornaments tend to be played in a short and soft manner, resulting in onsets characterised by a long attack with a slow energy rise (Gainza et al., 2005; Böck & Widmer, 2013).

1.1 Related work

There are relatively few studies concentrating on onset detection of flute signals within ITM. Gainza et al. (2004) and Kelleher et al. (2005) used instrument-optimised band-specific thresholds alongside a decision tree to determine note, cut or strike based on duration and pitch. Köküer et al. (2014) also analysed flute recordings, using an instrument-specific filterbank and a fundamental frequency estimation method using the YIN algorithm by De Cheveigné & Kawahara (2002) to minimise inaccuracies associated with octave doubling. More recently, Jančovič et al. (2015) presented a method for transcription of ITM flute recordings with ornamentation using hidden Markov models and Beauguitte et al. (2016) evaluated note tracking using a range of methods on a corpus of 30 tune recordings.

Onset detection techniques used in existing flute signal analysis have largely relied upon algorithms utilising signal processing, while state-of-the-art generalised onset detection methods use probabilistic modelling. Ali-MacLachlan et al. (2016) evaluated 11 methods that had previously performed well in the MIREX wind instrument class. `OnsetDetector` achieved the highest precision and F-measure scores. The use of bidirectional long short-term memory neural networks allows this model to learn the context of an onset based on past and future information, resulting in high performance in the context where soft onsets and features with small pitch deviations are coupled with other spurious events.

1.2 Motivation

The approach undertaken in this paper extends upon the work published in Ali-MacLachlan et al. (2016) in which onsets were detected through the use of the `OnsetDetector` system Eyben et al. (2010). Inter-onset segment classification was performed using an classification method based on a feed-forward neural network.

The `OnsetDetector` system was trained on a broad range of music making it effective at detecting a variety of instrument onsets. While note onset detection accuracy was very successful, ornament detection accuracies proved to be quite low by comparison. In an attempt to improve onset detection for ITM, we implemented an onset detection method based on a convolutional neural network (CNN) and trained this model specifically on ITM flute recordings. As we believe that the detection of ornament onsets to be context-dependent, we evaluate detection accuracy in relation to events that occur immediately before and after the detected events. This evaluation allows us to determine *where* onset detection errors occur and allows us to observe limitations in the detection of notes, cuts, strikes and breaths, in the context of traditional music being played authentically at a professional level.

The remainder of this paper is structured as follows: Section 2 outlines the proposed onset detection method and Section 3 presents our evaluation methodology and dataset. Section 4 presents the results of this evaluation and Section 5 presents conclusions and future work.

2. METHOD

Our onset detection method is based on a convolutional neural network (CNN) classification method. CNNs share weights by implementing the same function on sub-regions of the input. This enables CNNs to process a greater number of features at a lower computational requirement compared to other neural network architectures (i.e., multi-layer perceptron). High onset detection accuracies have been achieved by CNNs using larger input features (Schluter & Böck, 2014).

Figure 3 gives an overview of the implemented CNN architecture. The input features are first fed into two sets of convolutional and max pooling layers containing dropouts and batch normalisation. The output is then reshaped into a one-dimensional format before being run through a fully-connected layer and a softmax output layer.

2.1 Convolutional and max pooling layers

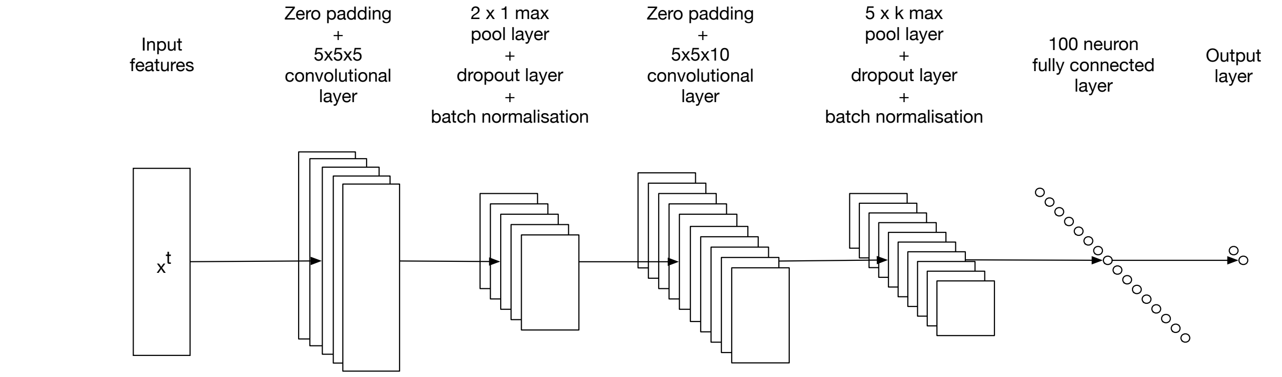
The output h of a two-dimensional convolutional layer with a rectified linear unit transfer function is calculated using:

$$h_{ij}^{f} = r\left(\sum_{l=0}^{L-1}\sum_{m=0}^{M-1} W_{ml}^{f} x_{(i+l)(j+m)} + b^{f}\right) \quad (1)$$

where x is the input features, W and b are the shared weights and bias and f is the feature map. L and M are the dimensions of the shared weight matrix and I and J are the output dimensions of that layer. The equation for the rectifier linear unit transfer function r is:

$$r(\phi) = max(0, \phi) \quad (2)$$

The output of the convolutional layer h was then processed using a max pooling layer which resulted in a $\frac{I}{a}$ by $\frac{J}{b}$ output where a and b are the dimensions of the sub-regions processed. A dropout layer (Srivastava et al., 2014) and batch normalisation (Ioffe & Szegedy, 2015) were then implemented.


Input features
Zero padding + 5x5x5 convolutional layer
2 x 1 max pool layer + dropout layer + batch normalisation
Zero padding + 5x5x10 convolutional layer
5 x k max pool layer + dropout layer + batch normalisation
100 neuron fully connected layer
Output layer
x^t

CNN5	512 x 5 x 1	512 x 5 x 5	256 x 5 x 5	256 x 5 x 10	51 x 5 x 10	100	2
CNN11	512 x 11 x 1	512 x 11 x 5	256 x 11 x 5	256 x 11 x 10	51 x 6 x 10	100	2
CNN21	512 x 21 x 1	512 x 21 x 5	256 x 21 x 5	256 x 21 x 10	51 x 7 x 10	100	2
CNN41	512 x 41 x 1	512 x 41 x 5	256 x 41 x 5	256 x 41 x 10	51 x 9 x 10	100	2
CNN101	512 x 101 x 1	512 x 101 x 5	256 x 101 x 5	256 x 101 x 10	51 x 11 x 10	100	2

Figure 3: Overview of the proposed implemented CNN system with different input feature sizes.

2.2 Fully-connected layer

A fully-connected layer consists of neurons which are linked to all of the neurons in previous and future layers. The output Y of a fully connected layer with a rectified linear unit transfer function is calculated using:

$$Y = r(W_c z + b_c) \tag{3}$$

where z is the input, W_c is the weight matrix and b_c is the bias. For the softmax output layer the rectified linear unit r transfer function is swapped for the softmax function which is calculated using:

$$softmax(\phi) = \frac{e^{\phi}}{\sum e^{\phi}} \tag{4}$$

2.3 Implementation

The CNN was implemented using the Tensorflow Python library (Abadi et al., 2016) with training data consisting of target activation functions created from ground truth annotations. A frame-based approach was taken where each frame is assigned 1 if it contains an onset or 0 if it does not.

2.4 Input features

Before processing by the CNN, the audio files must be segmented into frame-wise spectral features. An N sample length audio file was segmented into T frames using a Hanning window of γ samples ($\gamma = 1024$) and a hop size of $\frac{\gamma}{2}$. A frequency representation of each of the frames was then created using the discrete Fourier transform resulting in a $\frac{\gamma}{2}$ by T spectrogram. Various centred on the frame to be classified.

As classification is performed on the frame at the centre of the input features, a potentially crucial parameter is the number of input frames ψ. To determine the most efficient number of frames to use as the input for the CNN, five different values for ψ were used ($\psi = [5, 11, 21, 41, 101]$) creating the `CNN5`, `CNN11`, `CNN21`, `CNN41`, `CNN101` versions respectively.

2.5 Layer sizes

The layer sizes used for the different input features are indicated at the bottom of Figure 3. The size of all layers are consistent across systems apart from the second dimension k of the second max pooling layer. k is set to 1, 2, 3, 5 and 10 for the different input features sizes respectively.

2.6 Peak picking

The onsets must be temporally located from within the activation function Y output from the CNN. To calculate onset positions, the method from Southall et al. (2016) is used. A threshold τ is first determined using the mean across all frames and a constant λ:

$$\tau = \lambda \bar{Y} \tag{5}$$

The current frame t is determined to be an onset if its magnitude is greater than those of the surrounding two frames and above threshold τ.

$$O(t) = \begin{cases} 1, & y^t = max(y^{t-1:t+1}) \ \& \ y^t > \tau, \\ 0, & otherwise. \end{cases} \tag{6}$$

Finally, if an onset occurs within $25ms$ seconds of another then it is removed.

2.7 Training

The training data is divided into 1000 frame mini-batches consisting of a randomised combination of 100 frame re-

Player	Album(s)	Reels	Jigs	Polkas	Hornpipes
Harry Bradley	The First of May	8	4		4
Bernard Flaherty	Flute Players of Roscommon Vol.1	2			
John Kelly	Flute Players of Roscommon Vol.1		1		1
Josie McDermott	Darby's Farewell	2	2		2
Catherine McEvoy	Flute Players of Roscommon Vol.1, Traditional Flute Playing in the Sligo-Roscommon Style	4			
Matt Molloy	Matt Molloy, Heathery Breeze, Shadows on Stone	5	2		
Conal O'Grada	Cnoc Bui	13	1	10	
Seamus Tansey	Field Recordings	4			
Michael Tubridy	The Eagle's Whistle	2	9		
John Wynne	Flute Players of Roscommon Vol.1		3		

Table 1: Dataset recordings showing player, album source and tune type.

gions from the feature matrix. The Adam optimiser is used to train the neural networks with an initial learning rate of 0.003. Training is stopped when the validation set accuracy does not increased between iterations. To ensure training commences correctly, the weights and biases are initialised to random non-zero values between ± 1 with zero mean and standard deviation equal to one. The performance measure used is cross entropy and the dropout probability d is set to 0.25 during training.

3. EVALUATION

As the performance of the proposed method depends heavily on the accuracy of the chosen onset detection method, the aim of our first evaluation is to determine the quality of existing timing data. We then perform an evaluation of our onset detection method by comparing it against the most successful method found in Ali-MacLachlan et al. (2016).

3.1 Dataset

The corpus for analysis consists of 79 solo flute recordings by nine prominent traditional flute players. Four common types of traditional Irish tune are represented: *reels*, *jigs*, *hornpipes* and *polkas*. Individual players are discussed in Köküer et al. (2014) and players, tune type and recording sources are detailed in Table 1.

The dataset contains annotations for onset timing information and labels for notes, cuts, strikes and breaths, and is comprised of approximately 18,000 individual events. First notes of long rolls, short rolls and cranns were also identified and labelled.

3.2 Onset detection evaluation

The ground truth annotation process was completed using multiple tools as the project evolved (Köküer et al., 2014; Ali-MacLachlan et al., 2015) resulting in inconsistencies being found in onset placement and labelling. We therefore improved the quality of these annotations by comparing ground truth onsets against true positive and false negative onsets obtained using `OnsetDetector` (Eyben et al., 2010). Events outside a $50ms$ window of acceptance were evaluated by an experienced flute player, allowing events to be checked for onset accuracy. Patterns containing impossible sequences of events were identified and eliminated by checking each event in context with previous and subsequent events.

To obtain the results for the `OnsetDetector` system on the updated dataset all tracks were processed with the output onset times compared against the annotated ground truth. We assess the accuracy relating to the `OnsetDetector` method before and after annotation correction and the number of spectrogram frames used as input.

We then evaluate the `OnsetDetector` system against the implemented CNN systems the dataset is divided by tracks into a 70% training set (55 tracks), 15% validation set (12 tracks) and 15% test set (12 tracks). The training set is used to train the five versions of the CNN (`CNN5`, `CNN11`, `CNN21`, `CNN41`, and `CNN101`) onset detector using the different input feature sizes, the validation set is used to prevent over-fitting and the test set is used as the unseen test data. The `OnsetDetector` results for the 12 test tracks are compared to the results from the 5 CNN versions. F-measure, precision and recall are used as the evaluation metrics with onsets being accepted as true positives if they fall within 25ms of the ground truth annotations.

4. RESULTS

4.1 Onset detection results

	P	R	F
`OnsetDetector` Before annotation improvement	83.06	75.10	78.75
`OnsetDetector` After annotation correction	85.86	78.46	81.85
`CNN5`	87.06	84.71	85.73
`CNN11`	88.07	84.73	86.25
`CNN21`	88.82	**88.26**	**88.46**
`CNN41`	**88.84**	86.63	87.58
`CNN101`	88.72	86.21	87.32

Table 2: Precision (P), Recall (R) and F-measure (F) for `OnsetDetector` (Eyben et al., 2010) before and after annotation improvement, `CNN5`, `CNN11`, `CNN21`, `CNN41`, and `CNN101`.

			True Positives		
Label Code	Musical Pattern	Event Context	Onset Detector	CNN21	Total
111	note **note** note	*single notes*	1097	1124	1184
211	note **cut** note	*single cuts*	229	269	310
121	cut **note** note	*single cuts*	133	237	270
112	note **note** cut	*single cuts*	192	198	220
114	note **note** breath	*single notes*	96	99	106
411	note **breath** note	***single notes with breath***	21	53	88
311	note **strike** note	***single strike, end of roll***	55	42	76
122	cut **note** cut	*trill*	13	48	63
141	breath **note** note	*single notes*	55	56	61
131	strike **note** note	***single strike, end of roll***	16	33	57
123	cut **note** strike	*rolls*	14	33	36
261	note **cut** note	*start of long roll*	27	30	30
153	cut **note** strike	*start of short roll*	8	22	24
511	note **cut** note	*note before start of short roll*	18	21	23
612	note **note** cut	*note before start of long roll*	20	20	21
142	breath **note** cut	*breath before single cut*	19	20	20
241	breath **cut** note	*breath before single cut*	12	17	19
412	note **breath** cut	***breath before single cut***	3	11	19
115	note **note** cut	*two notes before start of short roll*	16	17	18
271	note **cut** note	*start of crann*	15	16	18
116	note **note** note	*two notes before start of long roll*	16	16	17
113	note **note** strike	*single strike*	14	13	15
117	note **note** note	*two notes before start of crann*	14	14	14
712	note **note** cut	*note before start of crann*	13	12	14
132	strike **note** cut	*cut after roll*	3	9	12

Table 3: Results comparing `OnsetDetector` and `CNN21` onset detectors for all event classes in the context of events happening prior and subsequent to the detected onset. Label codes of patterns with under 70% accuracy for `CNN21` shown in bold. Patterns with under 10 total onsets omitted.

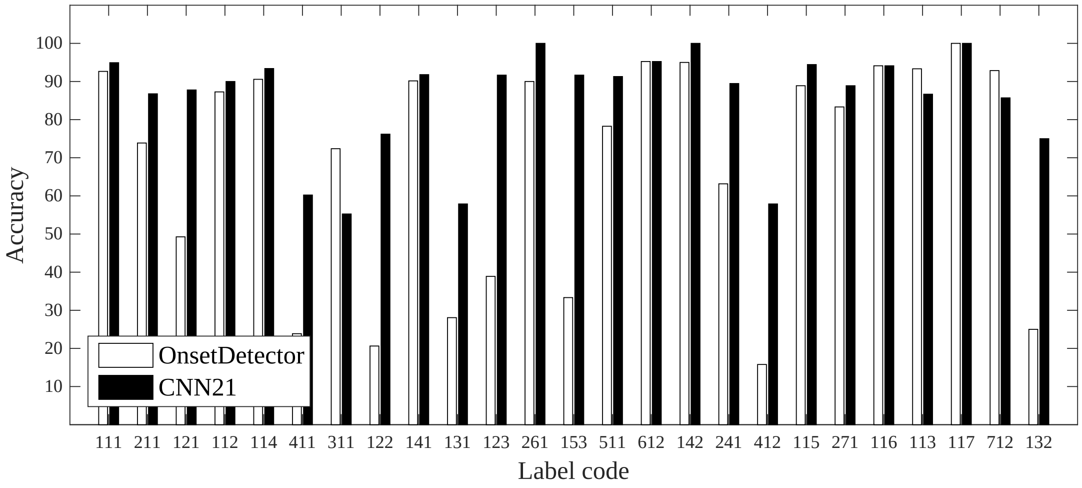


Figure 4: Accuracy of `OnsetDetector` and `CNN21` onset detectors for each event class above 10 onsets.

Table 2 presents the overall precision, recall and F-measure performance for the `OnsetDetector` and five CNN versions. The results indicate that all versions of the CNN achieve higher results than the `OnsetDetector`. The `CNN21`, which uses 10 spectrogram frames prior and subsequent to the middle frame achieves the highest recall and F-measure. The `CNN41` achieves a slightly higher precision than the `CNN21`, however achieves lower recall accuracy. The performance across the five CNN versions is fairly similar, illustrating that the moderate to higher values for the ψ parameter ($\psi = [21, 41, 101]$) are most appropriate for the task. The high performance of this approach is likely due to two factors. First, as CNNs are capable of processing large input feature sizes, they incorporate more context into the detection of a single frame. Second, as the CNNs are trained solely on traditional flute signals there is less variation in the represented classes, which has the potential of improving accuracy.

4.2 Note, cut and strike onset detection accuracy

Table 3 presents the onset detection results for each class of musical pattern with over 10 onsets in the test corpus of 12 tunes. The mean pattern precision across all classes was 79.22 for `CNN21` in comparison with 59.86 for `OnsetDetector`.

The classes consist of three event types where the central event is identified in bold. For example, label code 211 (*note **cut** note*) is a detected cut with a note before it and note after it, which exists within the event context of short and long roll or a single cut. The number of correctly detected onsets (true positives) is found as a percentage of the overall number of annotated onsets of that pattern. Label codes with an accuracy of less that 70% are shown in bold.

	Notes	Cuts	Strikes	Breaths
`OnsetDetector`	76.31	77.78	72.37	19.83
`CNN21`	89.57	91.29	55.26	59.06

Table 4: Accuracy of `OnsetDetector` and `CNN21` onset detectors for note, cut, strike and breath classes above 10 onsets.

As can be seen in Figure 4 and Table 3, low accuracies were found for strikes and notes following strikes. As a strike is played by momentarily tapping a finger over a tonehole, the pitch deviation is often much smaller than that of a cut and the event time is often shorter, making it more difficult to detect. Breaths are also difficult to detect in commercial recordings because it is usual to apply a generous amount of reverb effect at the mixing stage, resulting in a slow release masking a defined offset. Table 4 further illustrates inaccuracies in the detection of strikes and breaths by showing the accuracy for each single event class - note, cut, strike and breath. The note class also includes the notes at the start of ornaments such as long roll and crann and the cut class includes cuts at the start of short rolls.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an onset detection method based a convolutional neural network (CNN) and is trained solely on Irish flute recordings. The results from the evaluation show that this method outperformed the existing state-of-the-art generalised trained `OnsetDetector`. We have also improved the annotations of a ITM dataset by employing a process of automatic onset detection followed by manual correction as required. To evaluate the effectiveness of this approach, the top performing CNN version (`CNN21`) method is compared to the `OnsetDetector` by (Eyben et al., 2010), most successful method found in Ali-MacLachlan et al. (2016).

In future research, we aim to develop note and ornament classification methods with additional features and attempt other neural network architectures in order to capture trends that appear in time-series data. We plan to release a corpus of solo flute recordings that will allow a deeper study into differences in playing style, and to extend this corpus to include other instruments. We also plan to investigate the generality of the proposed system to other instruments characterised by soft onsets such as the tin whistle and fiddle. The dataset used in this paper will also be released shortly, alongside Köküer et al. (2017).

6. REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., & others (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*.

Ali-MacLachlan, I., Köküer, M., Athwal, C., & Jančovič, P. (2015). Towards the identification of Irish traditional flute players from commercial recordings. In *Proceedings of the 5th International Workshop on Folk Music Analysis*, Paris, France.

Ali-MacLachlan, I., Tomczak, M., Southall, C., & Hockman, J. (2016). Note, cut and strike detection for traditional Irish flute recordings. In *Proceedings of the 6th International Workshop on Folk Music Analysis*, Dublin, Ireland.

Beauguitte, P., Duggan, B., & Kelleher, J. (2016). A Corpus of Annotated Irish Traditional Dance Music Recordings: Design and Benchmark Evaluations.

Böck, S. & Widmer, G. (2013). Local Group Delay Based Vibrato and Tremolo Suppression for Onset Detection. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 589–594)., Curitiba, Brazil.

Boullier, D. (1998). *Exploring Irish Music and Dance*. Dublin, Ireland: O'Brien Press.

De Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

Eyben, F., Böck, S., Schuller, B., & Graves, A. (2010). Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 589–594)., Utrecht, Netherlands.

Gainza, M., Coyle, E., & Lawlor, B. (2004). Single-note ornaments transcription for the irish tin whistle based on onset detection. *Proc Digital Audio Effects (DAFX), Naples*.

Gainza, M., Coyle, E., & Lawlor, B. (2005). Onset detection using comb filters. New Paltz, New York, USA.

Hast, D. E. & Scott (2004). *Music in Ireland: Experiencing Music, Expressing Culture*. Oxford, UK: Oxford University Press.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, *abs/1502.03167*.

Jančovič, P., Köküer, M., & Baptiste, W. (2015). Automatic transcription of ornamented Irish traditional music using Hidden Markov Models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 756–762)., Malaga, Spain.

Keegan, N. (2010). The Parameters of Style in Irish Traditional Music. *Inbhear, Journal of Irish Music and Dance*, *1*(1), 63–96.

Kelleher, A., Fitzgerald, D., Gainza, M., Coyle, E., & Lawlor, B. (2005). Onset detection, music transcription and ornament detection for the traditional irish fiddle. In *Proceedings of the 118th AES Convention*, Barcelona, Spain.

Köküer, M., Ali-MacLachlan, I., Jančovič, P., & Athwal, C. (2014). Automated Detection of Single-Note Ornaments in Irish Traditional flute Playing. In *Proceedings of the 4th International Workshop on Folk Music Analysis*, Istanbul, Turkey.

Köküer, M., Ali-MacLachlan, Islah, Kearney, Daithi, & Jančovič, P. (2017). Curating and annotating a collection of traditional Irish recordings to facilitate stylistic analysis. *Special issue of the International Journal of Digital Libraries (IJDL) on Digital Libraries for Musicology, under review.*

Köküer, M., Kearney, D., Ali-MacLachlan, I., Jančovič, P., & Athwal, C. (2014). Towards the creation of digital library content to study aspects of style in Irish traditional music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, London.

Larsen, G. (2003). *The essential guide to Irish flute and tin whistle*. Pacific, Missouri, USA: Mel Bay Publications.

McCullough, L. E. (1977). Style in traditional Irish music. *Ethnomusicology*, *21*(1), 85–97.

Schluter, J. & Böck, S. (2014). Improved musical onset detection with convolutional neural networks. In *Acoustics, speech and signal processing (icassp), 2014 ieee international conference on*, (pp. 6979–6983). IEEE.

Southall, C., Stables, R., & Hockman, J. (2016). Automatic drum transcription using bi-directional recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 591–597)., New York City, United States.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.