PLAYER RECOGNITION FOR TRADITIONAL IRISH FLUTE RECORDINGS

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

Irish traditional music (ITM) is a form of folk music that developed alongside dancing over hundreds of years to become an integral part of Irish culture. The wooden flute is widely played in this tradition and mastery in performance is judged by personal stylistic interpretation. Automatic player recognition allows for musicological analysis in an environment where players are individuated based on their interpretation of a common set of melodies. This paper presents two player recognition methods based on convolutional neural networks (CNN). We implement two evaluation contexts for both methods, using a new *ITM-Flute-Style6* dataset alongside our existing *ITM-Flute-79* dataset. The results demonstrate that in both simplified and realistic scenarios, the proposed system is capable of high performance in recognising individual musicians playing melodies with individual stylistic traits that are idiomatic of the genre.

1. INTRODUCTION

Irish traditional music (ITM) is a solo and collective instrumental tradition with roots in social dance music (Vallely, 2011). Playing of the wooden simple system flute in ITM was historically linked to the west and northwest of Ireland (Williams, 2010) and traditional flute players are individuated based on their use of techniques such as ornamentation, phrasing and articulation (McCullough, 1977; Larsen, 2003; Hast & Scott, 2004; Keegan, 2010) alongside idiosyncratic timbral differences (Widholm et al., 2001; Ali-MacLachlan et al., 2013, 2015).

1.1 Related work

Musical genre classification is a widely studied area of music information retrieval (Fu et al., 2011) and an overview, including state of the art techniques, is presented by Sturm (2013). A subset of this field is musician recognition involving the definition of timbral, rhythmic and pitch content. Studies in flute acoustics have found that individual players produce markedly different timbres while changes in manufacturing material make very small spectral differences (Backus, 1964; Coltman, 1971; Widholm et al., 2001). Previous methods of player detection in ITM have used signal processing methods (Ali-MacLachlan et al., 2013, 2015).

Convolutional neural networks (CNN) have been successfully applied not only to image processing, but also to various audio analysis tasks, where the assumption is that auditory events can be recognised by analysing their time-frequency representations. To this end, CNNs provide multiple advantages to the task of musician recognition that other neural network models constitute impractical. The first benefit lies in the shared weights over the input that enable CNNs to process a greater number of features at a lower computational cost. This is achieved by applying the same function (filter) on sub-regions of the input images (spectrograms). This convolution operation is capable of feature translation that preserves the spatial information of the input, and can be used to learn musical features where the target musician's events can appear at any time or occupy any frequency range. CNNs have been implemented successfully with input features derived from spectrograms (Lee et al., 2009) and mel-frequency cepstral coefficients (MFCCs) representing timbre, tempo and key variations (Li et al., 2010). A CNN was trained to perform artist and genre recognition on the Million Song Dataset (Bertin-Mahieux et al., 2011) using segments related to note onsets and feature vectors containing timbre and chroma components (Dieleman et al., 2011). Lidy & Schindler (2016) used CNN to classify genre, mood and composer and achieved the highest results in MIREX 2016 using a 40-band Mel filter. Costa et al. (2017) reinforced the effectiveness of CNNs for music genre classification, performing analysis on three genres: Western, Latin and African music. Individual instrument classification is discussed in Park & Lee (2015) and as part of an ensemble in Han et al. (2017).

1.2 Motivation

In order to determine stylistic differences between players, we must first develop methods to recognise different players in audio signals. Earlier studies in player recognition for flute in ITM have relied upon existing collections of recordings where musicians do not play the same pieces of music (Ali-MacLachlan et al., 2015). We collect and evaluate recordings of six accomplished traditional flute players, all playing music from a predetermined corpus offering a range of typical modes and rhythms.

The use of deep learning, in particular CNN, is an important step in more accurate player identification. In order to identify flute players in ITM, we propose a CNN system in order to make use of their ability to efficiently process large datasets.

The remainder of this paper is structured as follows:

Section 2 details CNNs and their implementation. In Section 3 we discuss the evaluation strategies and the datasets used. Results of the studies into player recognition are presented in Section 4 and finally conclusions and further work are discussed in Section 5.

2. METHOD

We utilise a deep learning model to classify musician-specific features for the task of player recognition. More concretely, first the audio waveforms are pre-processed to create a desired signal representation in a form of MFCCs. Then the signal is split into 5-second segments that represent different timbral and rhythmic characteristics of each performer. Given these characteristics, our model aims to recognise the player of previously unseen audio segments. The next subsections discuss the single blocks of the system in more detail.

2.1 Feature Extraction

First, a downsampled 22.05 kHz 16-bit mono audio signal is split into five second segments. These audio segments offer enough information to capture rhythm patterns as well as result in a large number of observations. A magnitude spectrogram is then calculated using a 1024-sample window size and a resulting frame rate of 100 Hz. Finally, the frequency bins are transformed into 40 MFCCs in a frequency range from 32 Hz to 4,000 Hz.

2.2 Convolutional Neural Network

Figure 1 gives an overview of the implemented CNN architecture. The convolutional layers are constructed using two different building blocks that process the input features: Block A consists of a layer with 10 5x5 filters with 1x5 stride lengths and block B consists of a layer with 20 5x5 filters with 1x1 stride lengths; both are followed by max pooling ((2x2) and (2x5)), dropout layers (Srivastava et al., 2014) and batch normalisation (Ioffe & Szegedy, 2015). A fully connected layer with 100 neurons and a softmax output layer of size c (number of player classes) follows the convolutional blocks. This results in roughly 200,000 total parameters with slight variations depending on c.

2.3 Training

The Adam optimiser is used with a learning rate of 0.003 to train the model. Stochastic gradient descent is performed (batch size = 250) with the cross entropy loss function. Training is stopped when two criteria have been met: 1) 50 epochs have commenced and 2) validation set loss has not increased between epochs. The weights are initialised using a scaled uniform distribution (Sussillo, 2014) and biases are initialised to zero.

3. EVALUATION

Our proposed method of player recognition relies on the accuracy of recordings in representing the playing style of each musician. We implement four evaluation strategies, utilising 5-second audio segments from recordings to assess the ability of the proposed system to recognise different players.

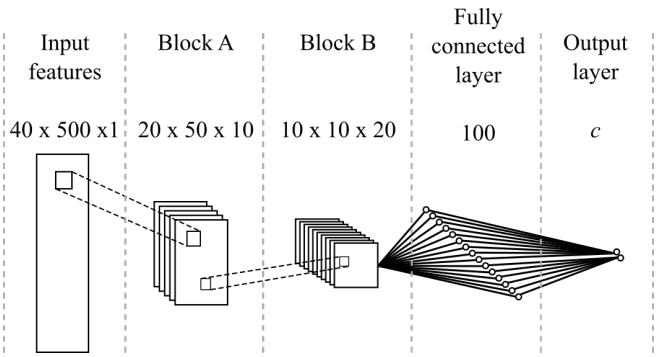


Figure 1: Overview of the proposed CNN. The input data flows between different network layers from left to right. The input data size at each computational block is presented above each layer.

3.1 Datasets

For the purposes of our evaluation we introduce a new *ITM-Flute-Style6* dataset and include a part of a previously used *ITM-Flute-99* dataset (Köküer et al., 2014; Ali-MacLachlan et al., 2015; Jančovič et al., 2015; Ali-MacLachlan et al., 2016, 2017).

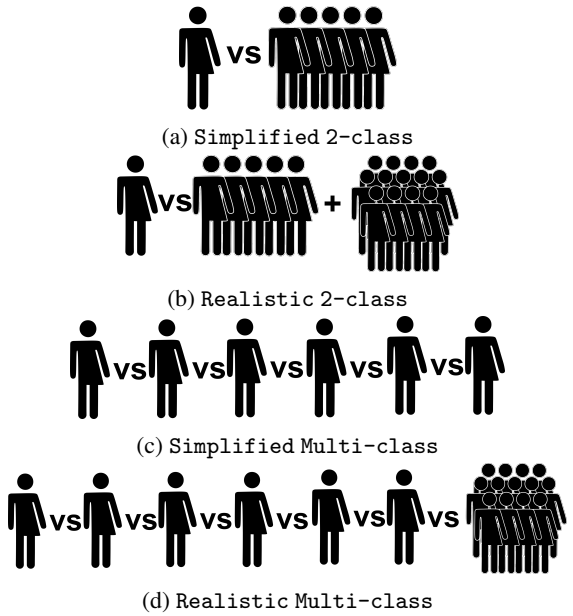
3.1.1 ITM-Flute-Style6

The new dataset consists of 28 recordings by each of 6 players (168 total) and is freely available on Github. [1] All tracks include only flute recordings with no accompaniment. The set covers a range of melodies or *tunes* that are common in the ITM community. The average duration of all tracks is approximately 43 seconds. The total duration of the dataset is 2 hours.

This dataset differs from the existing ITM flute datasets in that it targets multiple player traits and playing contexts that can substantiate further player style research. The presented tune types (i.e., *reels*, *jigs* and *hornpipes*) correspond to an informal online survey conducted among a group of experienced ITM players. The tune names can be seen in Table 1, where the last two represent individually chosen *wild* tracks by each player. The tune type category covers the three most popular tune types in ITM. Five categories are used to structure the dataset by: 1) player, 2) tune name, 3) tune type, 4) timed (*i.e.,* played to metronome) and 5) first or second repeat. All recorded flute players have substantial experience in playing and performing in the style of ITM. The timing category segregates the tracks into timed using a metronome, and untimed. All melodies were recorded twice in segue (first and second repeat) with and without metronome except wild tracks, which were only recorded without metronome.

The recordings were collected as 16-bit/44.1kHz WAV files using a Thomann MM-1 measurement microphone

[1] https://github.com/izzymaclachlan/datasets



(a) `Simplified 2-class`

(b) `Realistic 2-class`

(c) `Simplified Multi-class`

(d) `Realistic Multi-class`

Figure 2: Four evaluation strategies consisting of two player recognition approaches (`2-class` and `Multi-class`) and two contexts (`Simplified` and `Realistic`).

No.	Tune Title	Type	Scale	Ends on
1	Maids of Mount Cisco	Reel	G	Ray
2	The Banshee	Reel	G	Soh
3	Cooley's Reel	Reel	G	Lah
4	Banish Misfortune	Jig	G	Doh
5	Morrison's Jig	Jig	D	Ray
6	The Home Ruler	Hornpipe	D	Doh
7	Players choice 1	Wild	n/a	n/a
8	Players choice 2	Wild	n/a	n/a

Table 1: Corpus recorded by all players detailing tune type, scale and ending note.

connected to an Audient ID14 audio interface. The microphone was positioned above the middle of the flute in order to minimise wind noise caused by blowing.

3.1.2 ITM-Flute-99

ITM-Flute-99, includes 79 released recordings of 9 professional players detailed in Ali-MacLachlan et al. (2016). The remaining 20 recordings belong to a set of tutorial files by Larsen (2003) and were discarded due to the recordings being developed for teaching purposes rather than a true representation of the player. In our evaluation we treat the *ITM-Flute-99*, from now referred to as *ITM-Flute-79*, as representative of professionally played and recorded ITM flute performances.

3.1.3 Dataset experimental setup

The audio segments, described in section 2.1, are used to evaluate our player recognition accuracy. There are a total of 1885 segments comprising of 1438 from the *ITM-Flute-Style6* and 447 from *ITM-Flute-79*.

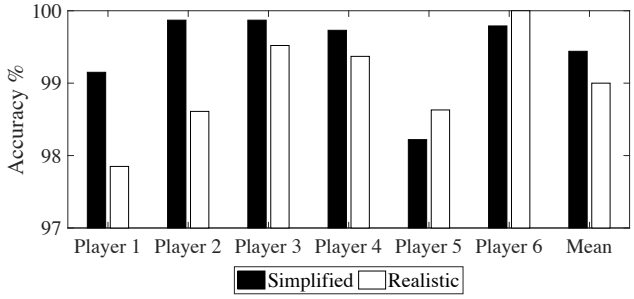


Figure 3: The `2-class` individual and mean player accuracies for both `Simplified` and `Realistic` contexts.

3.2 Evaluation Strategies

We implement four different evaluation strategies, consisting of two player recognition approaches (`2-class` and `Multi-class`) and two contexts (`Simplified` and `Realistic`) to test the system performance. An overview of the strategies is given in Figure 2. It is expected that the `Realistic` context will return lower accuracies as the larger dataset is representative of a wider range of player styles.

3.2.1 `Simplified 2-class`

In the first evaluation strategy, termed `Simplified 2-class` we use a `2-class` approach (using a softmax output layer with 2 neurons, *c=2*) to identify a single player from a mixed corpus. The first class (1) corresponds to the observed player and the second class (0) represents all other players. Due to there being only 6 players, the difference in total class observations should not cause significant bias during training. We test the `2-class` approach in a `Simplified` case using just the 6 player data from *ITM-Flute-Style6*. All recording variations of six tracks are used for training and all recording variations of the other two tracks are split evenly into validation and test sets. Four fold cross validation is performed so that each track appears in the test set. This is repeated 6 times with the player class corresponding to a different player each time.

		Pred	
		P	**O**
GT	**P**	99.7	0.6
	O	0.3	99.4

		Pred	
		P	**O**
GT	**P**	97.5	0.8
	O	2.5	99.2

Table 2: `2-class` confusion matrices where Pred is the predicted class, GT is the ground truth class, P is the player class and O the other class. The `Simplified` context is on the left and the `Realistic` context is on the right.

3.2.2 `Realistic 2-class`

In the second evaluation strategy, termed `Realistic 2-class`, we use the same `2-class` approach as Section

	1	2
M	99.7	100
N	100	100

	1	2
M	95.0	96.7
N	97.3	98.8

Table 3: `2-class` subgroup accuracies where M is metronome, N is no metronome, 1 is first repeat and 2 is second repeat. The `Simplified` context is on the left and the `Realistic` context is on the right.

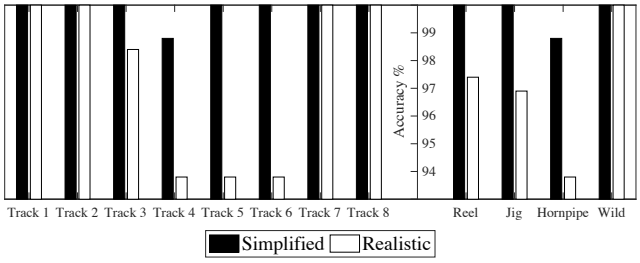


Figure 4: `2-class` results per track and the mean for the different track types. The `Simplified` context is on the left and the `Realistic` context is on the right.

3.2.1 but include *ITM-Flute-79* to create a more realistic evaluation. We expect that the inclusion of the additional data will reduce performance. All tracks from the added dataset are labelled as the other player class (0) and the dataset is divided by track into training, validation and testing respectively 75%, 12.5% and 12.5%.

3.2.3 `Simplified Multi-class`

In the third evaluation strategy, termed `Simplified Multi-class`, we aim to be able to classify an audio segment as one of multiple players. To do this a separate class is used for each of the 6 players (1,2,3,4,5,6) using a softmax output layer with 6 neurons (c=*6*). We then test the `Multi-class` approach using the same evaluation methodology as the `Simplified` context used in Section 3.2.1.

3.2.4 `Realistic Multi-class`

In the final evaluation strategy, termed `Realistic Multi-class` we test the `Multi-class` approach in a more realistic situation using the same two datasets and evaluation methodology from Section 3.2.2. However, all of audio segments from *ITM-Flute-79* are given their own new label (7).

4. RESULTS AND DISCUSSION

4.1 `2-class`

Figure 3 presents the results of each player and the mean across players for both `2-class` evaluation strategies (Figure 2(a) and 2(b)). As expected, a higher mean player accuracy is achieved in the `Simplified` context than the `Realistic` context. Player 5 achieves the lowest classification accuracy whereas player 6 achieves the highest. This could be due to player 6 having a much harder *tone* and higher harmonic energies whereas player 5 has a softer tone similar to the other four players.

A confusion matrix for the mean player results of the two `2-class` contexts are presented in Table 2. For the `Simplified` context (Figure 2(a)) there is approximately the same amount of misclassified player segments as there are misclassified other player segments. In the `Realistic` context there is a significantly higher amount of misclassified player segments and the greatest decrease in performance occurs for player 1 and player 2 (Figure 3), because these players show less individual stylistic traits like use of ornamentation or changes in timbre.

Table 3 presents the `2-class` dataset category distributions of correctly classified audio segments. For both contexts, first repeat with metronome (top left) achieves the lowest accuracy and second repeat without metronome (bottom right) achieves the highest accuracy. This makes sense as an artist is generally more reserved when they are trying to stay in time with a metronome or are playing a melody for the first time.

Figure 4 presents the percentages of the correctly classified player segments (same as Table 3) for each of the tracks. Also presented are the mean accuracies for the track types (Figure 4). The highest accuracies are achieved on the wild tracks. This could be due to the fact that the wild tracks are chosen by the players and suit their preferred playing technique. The lowest accuracies were achieved in the Jig and Hornpipe tracks (Tracks 4, 5 and 6) and they also see the largest decrease in accuracy between the `Simplified` and `Realistic` contexts. Reels are the most common tunes in ITM. As jigs and hornpipes are less common, musicians may be less comfortable playing this type of melody.

4.2 `Multi-class`

Table 4 presents confusion matrices for the `Multi-class` evaluation strategies. Again, as expected, a higher mean accuracy is achieved in the `Simplified` context than the `Realistic` context with 99.6% and 96.6% achieved respectively. While both approaches achieve a similar accuracy in the `Simplified` context the `Multi-class` approach achieves a lower accuracy than the `2-class` approach in the `Realistic` context. In order to recognise the work of a single player, the `2-class` approach is more accurate. As in the `2-class` evaluations, the highest overall accuracy is achieved when recognising player 6 and the lowest when recognising player 1. This is likely due to player 6 having a more individual style. The majority of the errors within the `Realistic` context are misclassified other players. Again, the few examples that are misclassified are of players with similar playing characteristics like timbre and amount of ornamentation.

In this study high accuracies are achieved suggesting that individual players can be recognised using spectral features, however the data consists of only a small number of flute players. It is expected that extending the *ITM-Flute-Style6* dataset would result in lower accuracies

		Prediction					
		1	**2**	**3**	**4**	**5**	**6**
Ground Truth	**1**	98.8	0	0	0	0	0
	2	0	100	0	0	0	0
	3	0	0	100	0	0.6	0
	4	0	0	0	100	0.9	0
	5	1.2	0	0	0	98.5	0
	6	0	0	0	0	0	100

		Prediction						
		1	**2**	**3**	**4**	**5**	**6**	**O**
Ground Truth	**1**	96	0.8	0	1.7	0	0	0
	2	0.7	97.8	0	0	0	0	0.8
	3	0.8	0	97.9	0	0.6	0	3.4
	4	0	0	0	98.3	0	0	2.3
	5	2.5	0	2.1	0	98.8	0	0.4
	6	0	0	0	0	0	100	1.3
	O	0	1.4	0	0	0.6	0	91.8

Table 4: `Multi-class` confusion matrices. `Simplified` context is on the left and the `Realistic` context is on the right.

as new players would be similar to those represented by existing recordings.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a flute player recognition method using a CNN trained on five-second solo excerpts. We evaluated the method using four strategies consisting of two approaches (`2-class` and `Multi-class`) and two contexts (`Simplified` and `Realistic`). For single player recognition, results from the evaluation show that the `2-class` method is more efficient. A player can be more easily recognised in second repeats and when not playing to a metronome. This suggests that players are less characteristic and more self-restricting when they play a tune for the first time or to a metronome. The highest accuracies are achieved when musicians play their own choice of melody (*i.e.,* wild tracks).

In future research, we aim to develop single note and ornament classification methods with additional features. We also aim to gain a deeper understanding of how the network is differentiating between stylistic features. We plan to implement other neural network architectures in order to compare the accuracy of different methods. In order to determine whether accuracies decrease with the addition of other players, we plan to extend the *ITM-Flute-Style6* dataset by recording additional flute players. We will follow the same methodology to record other traditional Irish instruments in order to compare stylistic traits across a range of instruments.

6. REFERENCES

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*, (pp. 13–17)., Paris, France.

Ali-MacLachlan, I., Köküer, M., Jančovič, P., Williams, I., & Athwal, C. (2013). Quantifying Timbral Variations in Traditional Irish Flute Playing. In *Proceedings of the 3rd International Workshop on Folk Music Analysis*, (pp. 7–13)., Amsterdam, Netherlands.

Ali-MacLachlan, I., Southall, C., Tomczak, M., & Hockman, J. (2017). Improved onset detection for traditional Irish flute recordings using convolutional neural networks. In *Proceedings of the 7th International Workshop on Folk Music Analysis.*, (pp. 73–79)., Malaga, Spain.

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.

Backus, J. (1964). Effect of wall material on the steady-state tone quality of woodwind instruments. *The Journal of the Acoustical Society of America*, *36*(10), 1881–1887.

Bertin-Mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, FL, USA.

Coltman, J. (1971). Effect of material on flute tone quality. *The Journal of the Acoustical Society of America*, *49*(2B), 520.

Costa, Y., Oliveira, L., & Silla, C. (2017). An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing*, *52*, 28–38.

Dieleman, S., Brakel, P., & Schrauwen, B. (2011). Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, (pp. 669–674)., Miami, FL, USA.

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia*, *13*(2), 303–319.

Han, Y., Kim, J., & Lee, K. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 208–221.

Hast, D. & Scott, S. (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. In *International conference on machine learning*, (pp. 448–456)., Lille, France.

Jančovič, P., Köküer, M., & Baptiste, W. (2015). Automatic transcription of ornamented Irish traditional music using Hidden Markov Models. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, (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.

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.

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

Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, (pp. 1096–1104)., Vancouver, Canada.

Li, T., Chan, A., & Chun, A. (2010). Automatic musical pattern feature extraction using convolutional neural network. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, (pp. 546–550)., Kowloon, Hong Kong.

Lidy, T. & Schindler, A. (2016). Parallel convolutional neural networks for music genre and mood classification. *MIREX*.

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

Park, T. & Lee, T. (2015). Musical instrument sound classification with deep convolutional neural network using feature fusion approach. *CoRR*, *abs/1512.07370*.

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.

Sturm, B. L. (2013). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, *41*(3), 371–406.

Sussillo, D. (2014). Random walks: Training very deep nonlinear feed-forward networks with smart initialization. *CoRR*, *abs/1412.6558*.

Vallely, F. (2011). *The Companion to Irish Traditional Music*. Cork, Ireland: Cork University Press.

Widholm, G., Linortner, R., Kausel, W., & Bertsch, M. (2001). Silver, gold, platinum-and the sound of the flute. In *Proc. Int. Symposium on Musical Acoustics*, (pp. 277–280)., Perugia, Italy.

Williams, S. (2010). *Irish Traditional Music*. Abingdon, Oxon: Routledge.