On the use of U-Net for dominant melody estimation in polyphonic music

Guillaume Doras
Sacem, Ircam
UMR STMS 9912, CNRS
Paris, France
guillaume.doras@sacem.fr

Philippe Esling
Ircam,
UMR STMS 9912, CNRS
Paris, France
philippe.esling@ircam.fr

Geoffroy Peeters
LTCI, Telecom ParisTech
University Paris-Saclay
Paris, France
geoffroy.peeters@telecom-paristech.fr

***Abstract*—Estimation of dominant melody in polyphonic music remains a difficult task, even though promising breakthroughs have been done recently with the introduction of the Harmonic CQT and the use of fully convolutional networks. In this paper, we build upon this idea and describe how U-Net – a neural network originally designed for medical image segmentation – can be used to estimate the dominant melody in polyphonic audio. We propose in particular the use of an original layer-by-layer sequential training method, and show that this method used along with careful training data conditioning improve the results compared to plain convolutional networks.**

***Index Terms*—dominant melody estimation, pitch estimation, HCQT, U-Net**

I. Introduction

Dominant melody or multi-pitches estimation in polyphonic music has long been seen as a difficult problem in Music Information Retrieval (MIR), both because of the inherent harmonic complexity of real-life music and because of the lack of annotated data available for training and evaluating.

Most of the successful approaches proposed so far for this task start by deriving a *pitch salience* from a spectral representation, and then apply some heuristics to it to estimate the dominant melody and/or the multiple pitches. Such heuristics are as varied as harmonic partials summation [1], pitch contour tracking [2], spectral smoothness enforcement [3], [4] or source-filter modeling [5], [6].

Recently, deep neural networks have been proposed to compute this pitch salience representation, using Recurrent Neural Networks (RNN) in [7], Convolutional Neural Networks (CNN) in [8], or a combination of both in [9]. The audio representation usually provided as input to the network is the Short Time Fourier Transform (STFT), but some authors have also used the raw waveform [10] or the Harmonic Constant-Q Transform (HCQT) [8].

In this paper, we propose the use of a *U-Net* architecture to estimate the dominant melody in polyphonic music. We propose a sequential method to train the U-Net using ground truth data at increasing resolutions, and show that this method improves performances compared to the usual training. We also compare the performances of the U-Net to those of the full CNN proposed in [8], and show that the U-Net architecture brings slight improvements over this previously proposed approach.

II. Related work

In this work, we build upon three main existing concepts: the HCQT data representation, the U-Net architecture and the curriculum learning paradigm.

A. Harmonic Constant-Q Transform (HCQT)

The HCQT, introduced in [8], is an elegant and astute representation of the audio signal in 3 dimensions (time, frequency, harmonic). It stacks along the third dimension several standard CQTs sharing the same frequency resolution and frequency range, but starting at different minimal frequency $h.f_{min}$, where f_{min} is the minimal frequency of interest and h is the harmonic index of each CQT. The harmonic components of the audio signal will thus be represented along the third axis of the HCQT and localized in the time-frequency domain across its first and second dimensions.

The alignment of harmonic series along the third dimension makes this representation particularly suitable for melody tracking, as it is can be directly processed by convolutional networks, whose 3-D filters can be trained to localize in the time and frequency plan the harmonic components in the melody of the input signal.

B. U-Net

U-Net was originally introduced in the context of image segmentation [11] for identifying and localizing high resolution details in medical images. It can be seen as a downsampling/upsampling model, where the downsampling part (the descending branch of the U) is learning representations of the input image at coarser resolutions by means of convolution and pooling layers, while the upsampling part (the ascending branch of the U) is learning to recreate representations at finer resolutions by means of convolution and transposed convolution layers [12]. The main difference with a convolutional Auto-Encoder is the introduction of *skip-connections* from the encoding levels to their counterpart decoding levels. These skip connections can be seen as a manner of providing an information context to the next reconstruction level, and have proven to help localization of features of interest.

The U-Net model has already been used in the context of MIR for sources separation [13]. We apply it here for dominant melody estimation, as we think that an analogy can

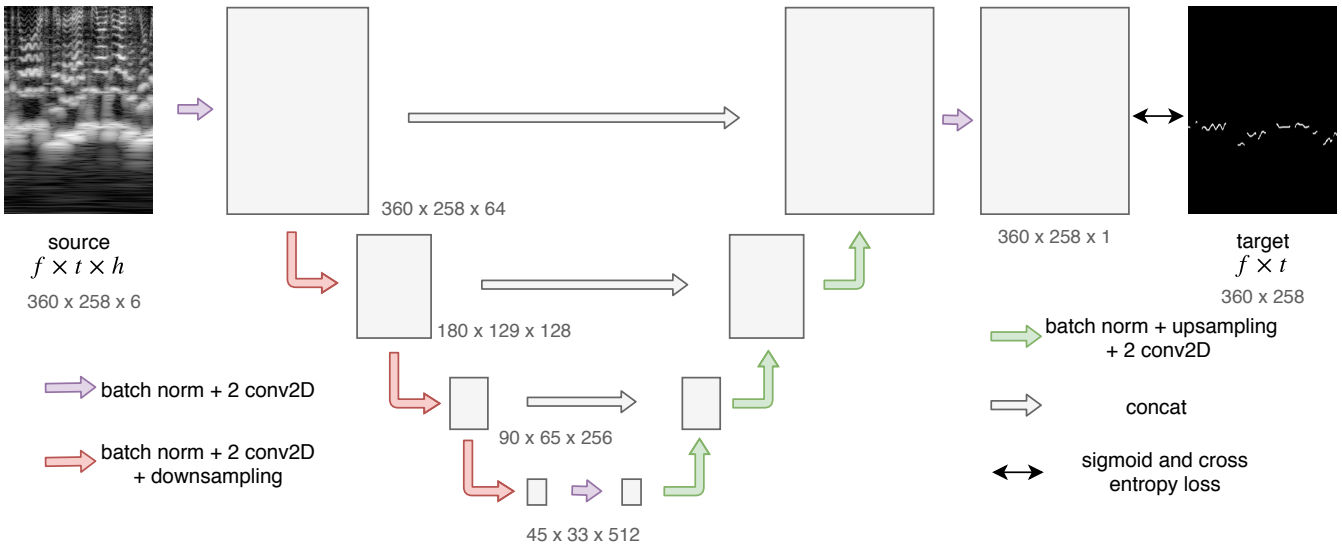

Fig. 1. U-Net model for dominant melody estimation.

be drawn between this problem and the image segmentation problem. Indeed, considering the HCQT as an image with h channels, contrasting and extracting the melody line from the background noise can be seen as a task similar to contrasting and extracting objects boundaries out of the rest of a natural image.

C. Curriculum learning

Curriculum learning was introduced as a continuation method, i.e. a strategy to minimize non-convex criteria [14], based on the intuition that a model – similarly to humans – could learn more efficiently if trained with successive training objectives of increasing difficulty, starting first with smooth objectives and gradually increasing the level of their complexity. This can be seen also as a sort of pre-training, which has proven to be beneficial [15].

The nature of the dominant melody estimation problem and the architecture of the U-Net are well suited for a curriculum learning approach: instead of training the model to deal with high resolution information directly, it is possible to prune parts of the network and train it repeatedly level by level. Successive trainings will start with coarse resolution information at the lowest level of the U, and continue with increasing resolutions while adding higher levels to the upsampling branch. This will be described more in details in section III.

III. Method

A. Model

The U-Net model used here is directly inspired by the seminal U-Net of [11], and is depicted in Fig. 1 with four levels.

On the *down-sampling branch*, each level consists of a batch normalization layer followed by two convolution layers with 3×3 kernels. Contrary to the original U-Net, padding is applied before convolutions ('same' convolution type), so that the time and frequency dimensions are maintained. Convolution layers are then followed by a max-pooling layer with a kernel of shape 2×2 and a stride of 2. The first level starts with 64 kernels, and the number of kernels is doubled at each level (i.e. the deeper level handles 512-depth tensors).

On the *up-sampling branch*, each level consists of a batch normalization layer followed by a transposed convolution layer with 2×2 kernels and a stride of 2 also, followed by two convolutional layers of 3×3 kernels also. The number of kernels is divided by 2 at each level.

At each resolution level, the output of the down-sampling branch is concatenated with the output of the up-sampling branch. For uneven dimensions on the down-sampling branch, the up-sampling will produce an even dimension. In this case, the supernumerary row or column is simply removed, so that data in each of the two branches has same shape and can be concatenated via the corresponding skip connection.

Finally, the output tensor is processed with a 1×1 kernel layer with sigmoid activation such that each time/frequency bin models a probability. The model is then trained to minimize the cross-entropy between the output probability and the the target ground truth normalized activations.

B. Data chunking

In order to process the full duration of songs and their corresponding dominant melody ground truth annotations, the data is split into chunks. Different durations of chunks have been tried, and preliminary experiments showed that a duration of 3 seconds is a good trade-off.

As padded convolutions are used, extra data might be added to the borders of the chunks on the frequency and the time axis. Additionally, removing supernumerary rows or columns at the borders when up-sampling might remove relevant information propagated from one layer to another.

We considered that zero padding on the frequency axis is acceptable, as lowest and highest frequencies of the HCQT are

very unlikely to be part of the melody. However, sides effects on time axis might not be negligible. To mitigate these effects, we have overlapped the beginning and the end of each chunk. The full duration melody estimation is reconstructed trimming each chunk's overlapping part and concatenating the remaining parts along the time axis. In practice, an overlap of 0.3 seconds at the beginning and at the end of each chunk appears adequate for 3 seconds chunks.

C. Curriculum training

We have investigated two different training methods: a classical end-to-end method, and a level-by-level method inspired by the curriculum learning paradigm.

In the level-by-level method, the up-sampling branch of the U-Net is initially pruned, except its lowest level. The ground truth target is downsampled with three pooling layers (that have no trainable parameters) to match the dimensionality of the lowest resolution level, as illustrated in Fig. 2. Only the down-sampling branch and the lowest level are then trained to minimize the cross entropy loss between the network output and the ground truth target at this coarsest resolution.

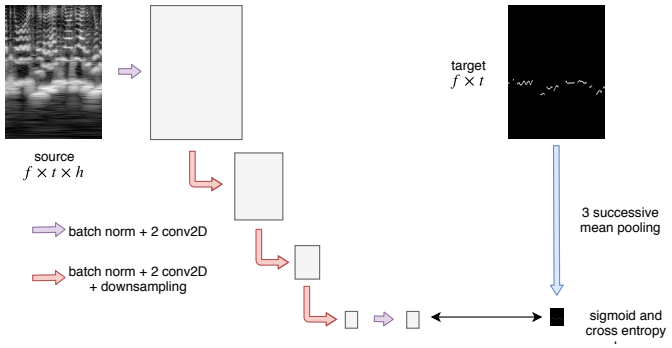

Fig. 2. Training of the lowest level of our Dominant Melody U-Net model

The next level layers and skip connections are then added to the partially trained network. The resulting network is trained to minimize the loss re-defined as the cross entropy between the new level output and the ground truth downsampled to the corresponding dimensionality.

Each next level is then subsequently added and the entire resulting network is trained reusing the weights of the lower levels. These successive partial trainings are repeated until the highest and finest resolution level is trained.

The main goal behind this strategy is to provide information about the ground truth target to the up-sampling branch as early as possible. Our assumption is that providing ground truth information at coarse resolution should help reconstruction at higher resolution levels.

IV. Dominant melody extraction experiment

A. Dataset

To train our networks, we have used the first release of the MedleyDB dataset [16][1], which provides the dominant melody and multi-pitch annotations for 108 songs of varied musical styles. We used the "melody2" annotations (see [17] for details) as dominant melody target for our networks.

Train/validation/test sets. Preliminary experiments have shown that different randomized train/validation/test sets splits could lead to very different results from one split to another. In order to obtain more robust results, we conducted a 10-folds cross-validation experiment. The 108 songs were divided into 10 folds containing 10 to 11 songs, using artist filtering (songs of the same artist must belong to the same fold). Each of the ten folds was used in turn as the test set. Another fold was randomly picked among the nine remaining ones to be used as the evaluation set, while the remaining eight folds were used together as the train set.

Baseline comparison. Because of this approach, we cannot directly compare our results to the ones published in [8]. In the following, we therefore consider as the baseline our own re-implementation of [8] applied to each of the 10-folds train/validation/test sets.

B. Configuration

For all experiments, we compute the HCQT as described in [8] with f_{min} = 32.7 Hz and 6 harmonics – $h \in \{0.5, 1, 2, 3, 4, 5\}$. Each CQT spans 6 octaves with a resolution of 60 bins per octave (5 bins per semi-tone), and has a frame duration of $\approx$ 11 ms. The implementation of the CQT was done with the Librosa library [18].

Training parameters. For training, we shuffled the chunks of the training set and then used a batch size of 16 chunks. We optimized the parameters using Adam [19] with a learning rate starting at 10^{-4} with a decaying factor of 0.94 per epoch. We applied early stopping if the loss on the validation set had not decreased after 1000 training steps.

From pitch saliency to dominant melody. The output of the networks (either the full CNN or U-Net) is a pitch salience representation. As in [8], we obtain the dominant melody simply keeping at each time frame the frequency with the maximum salience value. For the voicing/unvoicing decision at each time frame, we use a threshold whose value is chosen to optimize the Overall Accuracy score on the validation set. This threshold is then fixed and used on the test set before scores described below are computed.

C. Performance measures

To measure the performances of our system, we computed the melody Overall Accuracy (OA) along with the Raw Chroma Accuracy (RCA), Raw Pitch Accuracy (RPA) as well as the Voicing Recall (VR) and the Voicing False Alarm (VFA) scores as provided by the `mir_eval` toolbox [20].

V. Results

We compare here the three different systems: 1) our re-implementation of the fully convolutional baseline proposed in [8], 2) the U-Net trained end-to-end using chunks with temporal overlap, 3) the U-Net trained using curriculum training and chunks with temporal overlap.

[1]A newer and more accurate version has recently been released [17], but was not yet available during our experiments.

We show in Fig. 3 the distributions of the *mean* of metrics obtained by the three models for each of the 10 folds.

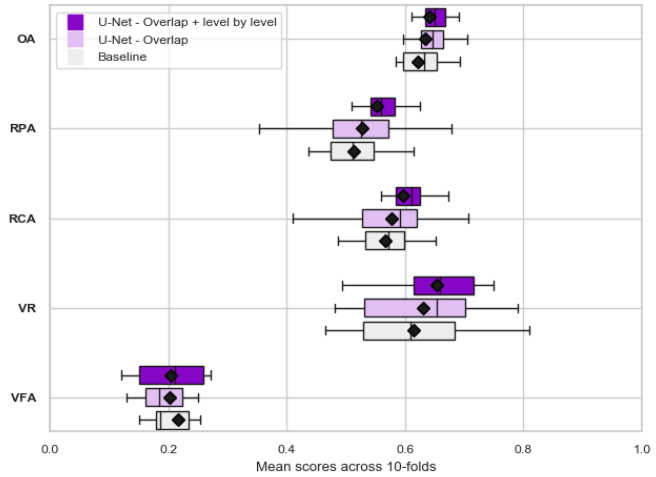

Fig. 3. 10-folds mean scores distributions obtained with `mir_eval` (OA=Overall Accuracy, RPA=Raw pitch accuracy, RCA=Row Chroma Accuracy, VR=Voicing Recall, VFA=Voicing False Alarm)

We see on Fig. 3 that the proposed U-Net provides some improvement for all scores compared to the baseline CNN.

Level-by-level training improvements. Now comparing the types of training used for U-Net, it appears that our proposed level-by-level training also provides further improvements on all scores, except for the Voicing False Alarm metric. Interestingly, the variance for these scores across the ten folds seems to be lower compared to the other models. This suggests that U-Net's generalization ability benefits from curriculum training, and that isolating and training lower levels first with coarse resolution data helps training of higher levels dealing with finer resolution data. The effect seems however less obvious on the Voicing False Alarm.

Voicing False Alarm. Despite the improved accuracies of U-Net, the Voicing False Alarm remains fairly high (around 20%). This is illustrated for a specific song in Fig. 4 where false voicing/unvoicing decisions are indeed often made: high values of pitch salience are present where no dominant melody is annotated. These voicing errors could be related to a discrepancy between the validation set (for which the voicing decision threshold has been optimized) and the test set. However, a visualization of the network outputs corresponding to empty chunks (i.e. chunks where no dominant melody is present) indicates that U-Net generally produces a non-empty output even when it should not. This suggests that conditioning the output with an extra voicing/unvoicing information could be beneficial, for instance with a dual loss [21].

Post hoc statistical significance analysis. All in all, the improvements of the mean scores observed in Fig. 3 between the different models remain fairly small. We have therefore conducted a Tukey's Honestly Significant Difference test (HSD) between each pair of models to assess the statistical significance of the observed improvements.

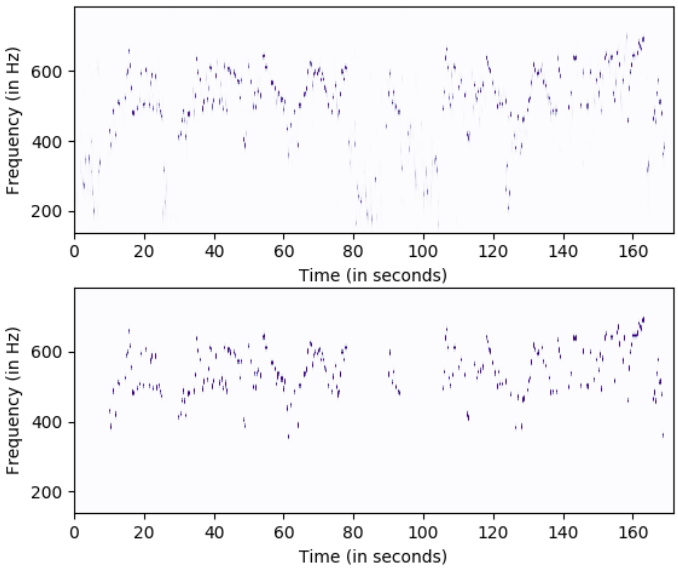

Fig. 4. [Top] Pitch salience output of U-Net trained level-by-level on overlapping chunks for the song of the test set "Don't Hear A Thing" by Brandon Webster. [Bottom] Corresponding MedleyDB's ground truth annotation.

The HSD test shows that the small differences between the mean scores of each fold are not large enough to reject the Null hypothesis, i.e. that the improvements observed on this dataset do not appear to be statistically significant enough to draw a definitive conclusion.

VI. Conclusion

In this paper, we have proposed to use the U-Net model for dominant melody estimation, and compared it with one of the current state-of-the-art models for this task, a fully convolutional network.

We have proposed to improve the performances of the U-Net model in two ways. Firstly, by overlapping training data to mitigate side-effects errors introduced by the padding and un-padding of its convolutions and de-convolutions layers. Secondly, by training U-Net with a curriculum training approach, starting with lower levels in isolation with coarse resolution data, and successively training higher levels with finer resolution data. We have shown that under these conditions, the U-Net provides a slight improvement over the full CNN. This improvement does however not appear to be statistically significant.

We however believe that the trend observed could be significant given larger amount of training examples. To improve performances of the proposed model, we therefore plan to use larger annotated datasets, such as Dali [22] or Lakh [23] datasets. We also plan to condition the network with a voicing/unvoicing information using a dual loss. Finally, we also want to continue exploring the idea that U-Net's higher levels can benefit from the knowledge of lower levels, and plan to introduce an attention mechanism between low and high resolution layers.

REFERENCES

[1] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes." in *ISMIR*, 2006, pp. 216–221.

[2] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[3] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[4] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[5] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[6] D. Basaran, S. Essid, and G. Peeters, "Main melody extraction with source-filter nmf and crnn," in *Proc. ISMIR*, 2018.

[7] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks." in *ICASSP*, 2012, pp. 121–124.

[8] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for f0 estimation in polyphonic music," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China*, 2017, pp. 23–27.

[9] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 5, pp. 927–939, 2016.

[10] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," *arXiv preprint arXiv:1802.06182*, 2018.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[13] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[14] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *ISMIR*, vol. 14, 2014, pp. 155–160.

[17] J. Salamon, R. M. Bittner, J. Bonada, J. J. Bosch, E. Gómez, and J. P. Bello, "An analysis/synthesis framework for automatic f0 annotation of multitrack datasets," in *Proceedings of the 18th ISMIR Conference*, 2017.

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[21] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *arXiv preprint arXiv:1710.11153*, 2017.

[22] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Dali: a large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm." in *19th International Society for Music Information Retrieval Conference*, ISMIR, Ed., 2018.

[23] C. Raffel, *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Columbia University, 2016.