2.3 From one staff to full page

To transcribe full-page music scores, a new interpretation of the obtained features map is considered. It is required to use an alternative reshape method from the output of the encoder block. In this paper, we apply a score unfolding reshape method. Instead of concatenating frame-wise elements along the height axis, as it is done for staff transcription, we reshape the image in the concatenation of all of its rows, obtaining then a $(h \times w, c)$ sequence. From a high-level perspective, this method can be understood as a staff concatenation process, as it is depicted in Fig. 2. This operation has to be done from top to bottom of the page, as it is crucial to correctly transcribe the music score.

Processing the feature maps this way prevents the aforementioned issues, as the score can be labeled as it is naturally read—from top to bottom and left to right—and the CTC method will not face the problem of collision of non-consecutive times, as now the used sequence does not have merged features from different staves.

This methodology, although theoretically a valid solution, presents some points that should be noted. The adaptation of the current music transcription models only needs the transcription of the full page for training, as the model directly outputs the music sequence from the input image. By avoiding the previously required object detection algorithm from the pipeline, now the system has to face two main challenges: (i) identifying all the staves in the score and (ii) transcribing them into an ordered sequence. The second challenge is solved by the reshape method, as we force the model to align and read all the staves in a specific order. In this case, the CTC *blank* token—which is an additional element introduced in the notation vocabulary to indicate time step separations—denotes both music element separations and staff breaks, since the single-staff-like produced sequence identifies the first symbol of the next staff as a consecutive timestep to the last symbol of the previous one. Challenge (i), however, is somewhat relegated to the network learning. The hypothesis is that, by not modifying the sequence structure in the decoder block, this alignment can be learned and mapped by the encoder. This hypothesis needs to be validated by the performance of the proposed models during experimentation.

2.4 Further considerations

This paper evolves the already established OMR staff recognition models by implementing a learned music score unfolding. Theoretically, this method is still a single-staff transcription system. All the staves of the pages can be understood as a single long staff that, due to physical constraints of paper, had to be divided into several staves. This methodology implicitly learns to reconstruct this original interpretation and transcribe the resulting long single-staff image in one step. The alignment process between the different staves is learned by the network during training.

It is important to note that, at this point, the proposed method is suitable only for monophonic staves, as polyphony requires additional information and processing to be transcribed holistically. Vocal music scores—such as those written in mensural notation—can benefit from this approach because polyphony is usually written through a series of independent monophonic voices. For this reason, the experiments will be carried out with early music scores written in mensural notation, as detailed in Section 3.2.

3. EXPERIMENTAL SETUP

In this section, we describe the proposed models to address full-page OMR, present the used corpora during experimentation, and define the metrics used to evaluate the performance of our proposals [1] .

3.1 Models

As mentioned throughout this paper, we are adapting state-of-the-art OMR models to full-page transcription. However, two additional variations have been included in the proposed model to extend the study on this topic. All the presented models have a fully convolutional block, which acts as an encoder of the input image features. This network is composed of stacked convolutional layers, which end up producing a feature map of size $(h/32, w/8, c, b)$, h and w being the height and the width of the input image, c the filters in the last convolutional layer, and b the batch size. The downscale of the original image size is produced by pooling operators. Then, the decoding architectures proposed for processing the sequence obtained after the reshaping procedure are described below.

3.1.1 Recurrent Neural Network

We follow the implementation of the original CRNN-CTC staff transcription model from [10], where the reshaped feature map is fed into a Bidirectional LSTM (BLSTM) and linearly projected onto the music notation dictionary. Specifically, we implemented a BLSTM with 256 units, whose output matches the output space of the fully convolutional network.

3.1.2 The Transformer

The base model of this work uses RNNs to process the reshaped feature map as a sequence. However, a recent recurrent-free model has gained popularity in the Natural Language Processing (NLP) field: the Transformer [18]. This model replaces the RNN architecture by implementing sequence modeling through attention mechanisms and position learning. This model solves some common issues that RNNs have—such as processing long sequences, training time effort, and contextual information retrieval—at the cost of needing more data to converge. As we have observed in the reshape step, the model would have to process significantly long sequences in one step, something that can have a negative impact on the performance of RNNs. For this reason, the first alternative model implemented in this paper replaces the recurrent layer of the CRNN model with a Transformer encoder (referred to as

[1] Source code for the implementation of the presented models can be found in `https://github.com/antoniorv6/ismir_fpomr.git`

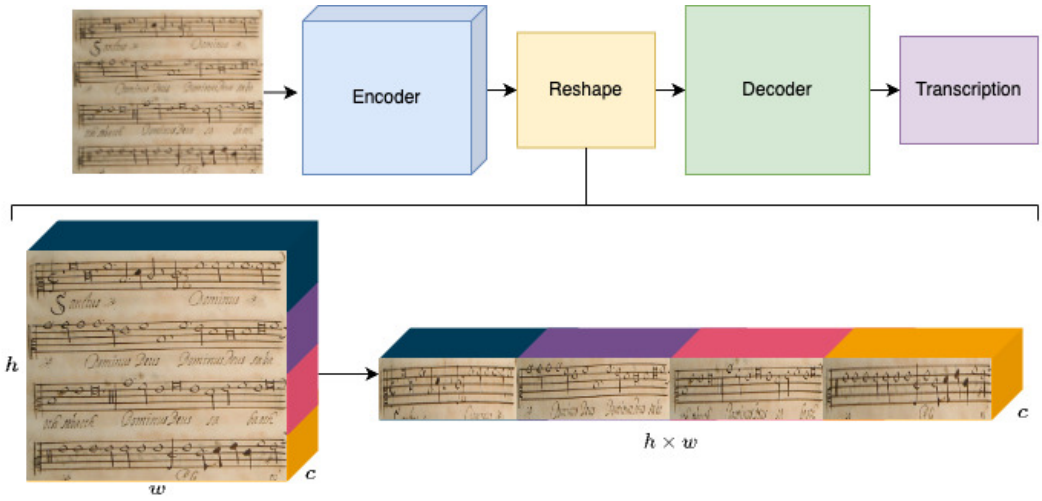


Figure 2: Graphic visualization of the reshape method to adapt current systems to full-page transcription. The reshape module learns how to separate and concatenate the staves on a page in a single line. Note that the original image has been used in the reshape for clarity of explanation, but the alignment is done in the *feature space* extracted by the encoder.

.

CNNT hereafter). In particular, we implemented one encoder layer with an embedding size of 512, a feed-forward dimension of 1024, and 8 attention heads.

3.1.3 Sequence-processing-free module

One of the challenges the model has to face during the score recognition is the alignment of the staves extracted from a 2D image to build a 1D sequence. This leads to the question of how a sequence processing module could impact learning this specific task while performing backpropagation. In analogous text recognition models, like [19], the solution lies in preserving the prediction space in 2 dimensions, applying backpropagation directly to the retrieved feature map before being reshaped. We have implemented this model to analyze the impact of the sequence processing module on the score page transcription. It is also based on fully convolutional layers, so it will be referred to as FCN in the results section.

3.2 Corpora

Two mensural-notation music datasets with different characteristics in engraving style were used, in order to represent the different challenges that the model can face in these real-case scenarios.

The first corpus is "Il Lauro Secco" [20] (denoted as SEILS), which corresponds to an anthology of 150 typeset printed images of the 16th-century Italian madrigals.

The second corpus is the CAPITAN dataset [10], which contains a complete ninety-six pages manuscript from the 17th-century containing a handwritten *missa*.

Specific details about the used corpora can be found in Table 1, and Fig. 3 depicts examples of these documents.

3.3 Data augmentation

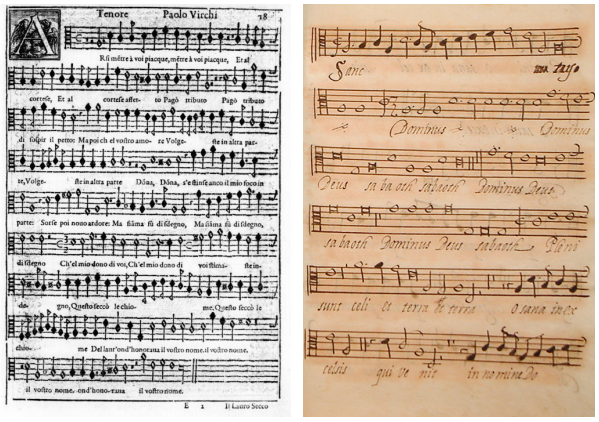
One constraint that our proposed model can find observing Table 1 is the scarcity of data, as these kind of early music documents usually have few pages completely labeled. To address this potential issue, we applied a data augmentation process to increase the number of samples per corpus. This procedure is composed of several image distortion operations, such as reduction, erosion, dilation, or perspective modifications. These distortions are randomly applied for each batch sample, allowing us to obtain massively extended corpora that also have an added variability for the samples. The SEILS corpus is increased up to 29000 pages and the CAPITAN dataset to 24000.

Table 1: Details of the corpora regarding the pages' features, such as sizes in pixels, number of samples and staves per page, number of symbols present and unique symbols per dataset (*vocabulary*).

	SEILS	**CAPITAN**
Num pages	150	123
Max page size	1200×813	1593×2126
Min page size	1200×813	1100×780
Avg staves per page	4	10
Max staves per page	9	12
Min staves per page	1	2
Avg symbols per page	222	136
Max symbols per page	331	220
Min symbols per page	110	23
Unique symbols	183	321

3.4 Metrics

Currently, there are no OMR-specific metrics to evaluate the performance of the transcription systems. In this paper, we resort to the Symbol Error Rate (SER), which is the most commonly used metric in the OMR literature for end-



(a) SEILS (b) CAPITAN

Figure 3: Music page examples from the used corpora.

to-end approaches. This metric is computed as

$$\text{SER}\,(\%) = 100\,\frac{\text{ED}(\hat{S}, S)}{|S|}\,,$$

where ED is the edit distance between the transcription hypothesis $\hat{S}$ and its corresponding ground truth, S. $|S|$ is the length (in tokens) of S. We chose this metric because it represents accurately the recognition performance of the model and correlates with the effort that a user would have to invest to manually correct the output sequence.

The datasets have been split into fixed partitions, where 60% of the samples have been used for training, 20% have been used for validation, and 20% for testing.

4. RESULTS

Table 2 shows the results obtained by the studied methodology on the test set for each corpus. Results reported by Castellanos et al. [12] are included to establish a reference value from a state-of-the-art algorithm based on a standard OMR pipeline (the one depicted in Fig. 1-top). Note that this reference model is trained under more favorable conditions, as it addresses the full-page transcription in two separate tasks, which have specific data for training each. However, it requires much more labeling work to build the training sets than our segmentation-free implementation. Therefore, it should be understood only as a reference and not as a competing approach.

The results obtained show that the extended models were able to transcribe full-page scores with fair SER values, below 30% except for the FCN without data augmentation in the CAPITAN dataset. These error values also scale depending on the engraving complexity of each corpus, being the printed documents (SEILS) easier to transcribe than the handwritten ones (CAPITAN). The model that reported the best results was the combination of the convolutional network with the RNN decoder (CRNN), which obtained an error rate of 4.3% in the SEILS corpus and 15.5% in CAPITAN.

The models that use sequence processing decoders (CRNN and CNNT) performed better than the single FCN network. This was expected to some extent, as these architectures exploit and optimize sequential information to improve their performance, which is a considerable advantage over the FCN approach. In fact, they also seem to bring some robustness to the model against the scarcity of training data, if we compare their results with the high error rate of the FCN when no data augmentation was applied to the CAPITAN corpus.

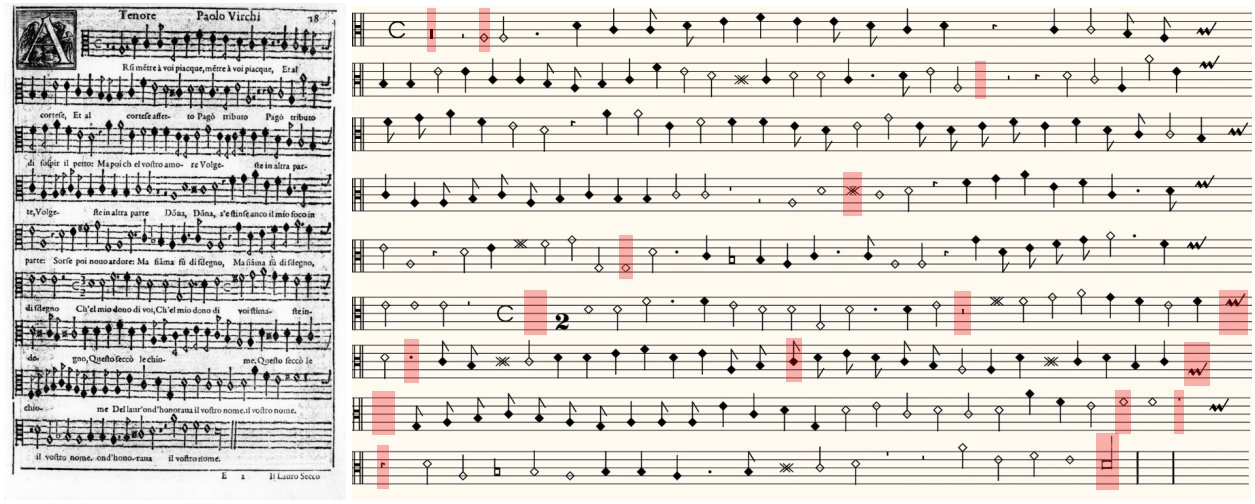
Table 2: Test SER (%) obtained for the studied models. Castellanos et al.'s work is included in the last row as a reference value to observe how current OMR pipelines work to transcribe the used corpora in this paper.

Model	Augmentation	SEILS	CAPITAN
CRNN	-	6.3	26.6
	✓	**4.3**	**15.5**
CNNT	-	12.9	28.4
	✓	7.2	18.2
FCN	-	23.3	89.5
	✓	13.3	22.5
Staff-based [12]	-	3.6	10.8

Continuing the data dependency analysis, it can be observed that applying data augmentation, significantly improved the overall performance of all models. The most notable case of this dependency was found in the FCN network on the CAPITAN corpus, where overfitting issues seemed to be solved with the augmented database. For the other models, improvements were reported as well, reducing the SER by approximately 30%–40%. In other words, the models are able to work with few samples, but they require a considerable amount of data to obtain their best results.

Comparing the two sequence processing decoders, we observe that RNNs performed better than CNN Transformers in all cases. The reason for this is aligned with the results obtained in works that explore the use of these models on document transcription [22]. Looking at the Table 1, we observe that both models contain, on average, short sequences. Transformers, by replacing recurrence with self-attention and position encoding, improved computation time and accuracy at the cost of more data needed to converge. However, the Transformers literature has reported relevant improvements with long sequences, containing approximately 512 tokens, with which the RNNs have convergence problems. In this case, the Transformer model is in a disadvantageous scenario, where few data samples are available to train and the output sequences are relatively short, which is a much better scenario for RNN-based decoders.

For the sake of visualization, Fig. 4 presents the results obtained in a test set page from the SEILS dataset by the best model (CRNN). As can be seen, the system produces a good transcription, in which most of the symbols are correctly labeled and aligned within their corresponding



(a) Original page.

(b) Transcribed prediction with CRNN, rendered with the Verovio toolkit [21].

Figure 4: Visualization of the transcription produced by the CRNN model in a music page from the SEILS dataset. Errors are highlighted. Note that the output is actually produced on a single line, but a multi-line representation of the score has been reconstructed to facilitate comparison with the original document. In this particular case, the obtained SER is 6.1% (higher than the average on this corpus).

staves. If we analyze the produced errors, most of them are subtle—such as vertical position misplacement— and can be easily corrected with score editing software. This happens often with narrow symbols—such as rests—whose manual annotation would also have been difficult to perform.

5. CONCLUSION

In this paper, we present a first segmentation-free end-to-end approach for full-page OMR. This method is trained with weakly-annotated data: it only requires a set of page images with their corresponding transcription, in contrast to current state-of-the-art full-page OMR pipelines that require spatial information—such as bounding boxes or pixel-wise regions. Our methodology extends the current staff-level end-to-end systems to full-page transcription by applying a concatenation step that learns how to process the two-dimensional sequence document.

We evaluated three variants of this model with two early music collections written in mensural notation, which have been used in many other works as a benchmark for OMR. The reported results showed that the proposed system produces competitive results for full-page transcriptions. Although precision is slightly lower than a multi-stage OMR pipeline, the proposed segmentation-free approach stands as an interesting alternative for those models. The model provides a favorable trade-off between the cost of labeling and the system's accuracy.

Several avenues for future research arise from this work. First, it only covers the extent of historic vocal music in mensural notation, where only monophonic staves are found. In fact, the method can handle any page representing monophonic staves but, in modern music, the coverage is more restricted, since polyphony is much more common. This scenario could be an interesting case to analyze alternative architectures for OMR, such as attention-based systems [23] or non-CTC-based image-to-sequence architectures [24], which do not have to be constrained by the specific layout.

Another aspect to consider in the future is the need for data. Although one of the goals of this work is to reduce the labeling effort, the models still require a large number of fully annotated pages. We believe that further research needs to be done to study both transfer learning and self-supervised learning approaches to address this issue.

6. ACKNOWLEDGEMENTS

This paper is part of the project MultiScore (PID2020-118447RA-I00), funded by MCIN/AEI/10.13039/501100011033. The first author is supported by grant ACIF/2021/356 from "Programa I+D+i de la Generalitat Valenciana". Third author was supported with a 2021 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. The Foundation takes no responsibility for the opinions, statements and contents of this paper, which are entirely the responsibility of its authors.

7. REFERENCES

[1] A. Hankinson, P. Roland, and I. Fujinaga, "The music encoding initiative as a document-encoding framework." in *Proc. of the 12th Int. Society for Music Information Retrieval Conference*, 2011, pp. 293–298.

[2] D. Huron, "Humdrum and kern: Selective feature encoding," *Beyond MIDI*, 1997.

[3] J. Calvo-Zaragoza, J. Hajič Jr., and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys*, vol. 53, no. 4, Jul. 2020.

[4] J. A. Burgoyne, L. Pugin, G. Eustace, and I. Fujinaga, "A comparative survey of image binarisation algorithms for optical recognition on degraded musical sources," in *Proc. of the 8th Int. Conf. on Music Information Retrieval*, pp. 509–512.

[5] J. dos Santos Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. P. da Costa, "Staff detection with stable paths," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 1134–1139, 2009.

[6] A. Pacha and H. Eidenberger, "Towards a universal music symbol classifier," in *Proc. of 14th Int. Conf. on Document Analysis and Recognition*, 2017, pp. 35–36.

[7] A. Fornés, J. Lladós, and G. Sánchez, "Old handwritten musical symbol classification by a dynamic time warping based method," in *Graphics Recognition. Recent Advances and New Opportunities*, W. Liu, J. Lladós, and J.-M. Ogier, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 51–60.

[8] F. Rossant and I. Bloch, "Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 081541, 2006.

[9] A. Pacha, J. Hajič, and J. Calvo-Zaragoza, "A baseline for general music object detection with deep learning," *Applied Sciences*, vol. 8, no. 9, p. 1488, 2018.

[10] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Handwritten music recognition for mensural notation with convolutional recurrent neural networks," *Pattern Recognition Letters*, vol. 128, pp. 115–121, 2019.

[11] A. Baró, C. Badal, and A. Fornés, "Handwritten historical music recognition by sequence-to-sequence with attention mechanism," in *Proc. of the 17th Int. Conf. on Frontiers in Handwriting Recognition*. IEEE, 2020, pp. 205–210.

[12] F. J. Castellanos, J. Calvo-Zaragoza, and J. M. Iñesta, "A neural approach for full-page optical music recognition of mensural documents," in *Proc. of the 21th Int. Society for Music Information Retrieval Conference*, 2020, pp. 12–16.

[13] M. Kletz and A. Pacha, "Detecting staves and measures in music scores with deep learning," in *Proc. of the 3rd Int. Workshop on Reading Music Systems*, J. Calvo-Zaragoza and A. Pacha, Eds., Alicante, Spain, 2021, pp. 8–12.

[14] M. Alfaro-Contreras, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "OMR-assisted transcription: A case study with early prints," in *Proc. of the 22nd Int. Society for Music Information Retrieval Conference*, 2021, pp. 35–41.

[15] A. Ríos-Vila, D. Rizo, and J. Calvo-Zaragoza, "Complete optical music recognition via agnostic transcription and machine translation," in *Proc. of the 16th Int. Conf. on Document Analysis and Recognition*, J. Lladós, D. Lopresti, and S. Uchida, Eds., 2021, pp. 661–675.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the 23rd Int. Conf. on Machine Learning*, 2006, p. 369–376.

[17] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, "DeepScores-A Dataset for Segmentation, Detection and Classification of Tiny Objects," in *Proc. of the 24th Int. Conf. on Pattern Recognition*, 2018, pp. 3704–3709.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[19] D. Coquenet, C. Chatelain, and T. Paquet, "Span: A simple predict & align network for handwritten paragraph recognition," in *Proc. of the 16th Int. Conf. on Document Analysis and Recognition*, 2021, pp. 70–84.

[20] E. Parada-Cabaleiro, A. Batliner, and B. W. Schuller, "A Diplomatic Edition of Il Lauro Secco: Ground Truth for OMR of White Mensural Notation," in *Proc. of the 20th Int. Society for Music Information Retrieval Conference*, 2019, pp. 557–564.

[21] L. Pugin, R. Zitellini, and P. Roland, "Verovio: A library for engraving MEI music notation into SVG," in *Proc. of the 15th Int. Society for Music Information Retrieval Conference*, 2014, pp. 107–112.

[22] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-line recognition," *Pattern Recognition*, vol. 129, p. 108766, 2022.

[23] D. Coquenet, C. Chatelain, and T. Paquet, "End-to-end handwritten paragraph text recognition using a vertical attention network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] S. S. Singh and S. Karayev, "Full page handwriting recognition via image to sequence extraction," in *Proc. of the 16th Int. Conf. on Document Analysis and Recognition*, 2021, pp. 55–69.

MEL SPECTROGRAM INVERSION WITH STABLE PITCH

Bruno Di Giorgi* Apple bdigiorgi@apple.com

Mark Levy* Apple mark_levy@apple.com

Richard Sharp Apple richard_sharp@apple.com

ABSTRACT

Vocoders are models capable of transforming a low-dimensional spectral representation of an audio signal, typically the mel spectrogram, to a waveform. Modern speech generation pipelines use a vocoder as their final component. Recent vocoder models developed for speech achieve a high degree of realism, such that it is natural to wonder how they would perform on music signals. Compared to speech, the heterogeneity and structure of the musical sound texture offers new challenges. In this work we focus on one specific artifact that some vocoder models designed for speech tend to exhibit when applied to music: the perceived instability of pitch when synthesizing sustained notes. We argue that the characteristic sound of this artifact is due to the lack of horizontal phase coherence, which is often the result of using a time-domain target space with a model that is invariant to time-shifts, such as a convolutional neural network.

We propose a new vocoder model that is specifically designed for music. Key to improving the pitch stability is the choice of a shift-invariant target space that consists of the magnitude spectrum and the phase gradient. We discuss the reasons that inspired us to re-formulate the vocoder task, outline a working example, and evaluate it on musical signals[1]. Our method results in 60% and 10% improved reconstruction of sustained notes and chords with respect to existing models, using a novel harmonic error metric.

1. INTRODUCTION

In modern speech synthesis pipelines a first model generates a low-dimensional audio representation, usually the mel spectrogram, from text; and a second model, named Vocoder, transforms the mel spectrogram to an audio waveform. Theoretically, vocoders designed for speech could be directly applied to musical signals; however closer inspection reveals features and constraints that are exclusive to the music domain. For example, unlike speech, music signals can be polyphonic and contain longer sustained notes whose pitch precision and stability is essential.

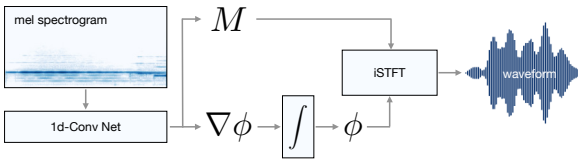


Figure 1: Our proposed model for mel spectrogram inversion. A one dimensional CNN estimates the magnitude and the phase gradient from the mel spectrogram. The phase gradient is then integrated to estimate the phase spectrum and finally audio is obtained via the inverse STFT.

The stability of a sustained pitched note manifests in the time-domain audio signal as the steady repetition of a periodic waveform. Periodic patterns are by definition not shift-invariant, except for shifts of an integer number of periods, therefore they require some form of auto-regression in order to be reproduced accurately. As expected, time-domain vocoders using shift invariant architectures [1, 2], despite other advantages such as generation efficiency, produce jitters that are perceived as pitch and timbre instability. For this reason, other time-domain generative models for audio include an autoregressive mechanism in the neural architecture [3–5]. In practice, time-domain models are required to learn all possible shifts of periodic patterns, a space that increases exponentially for polyphonic music, and how to create smooth sequences of these patterns.

Inspired by a recent generative model for single notes [6], we propose a new vocoder model for music (Fig. 1), where the target of the neural network is an intermediate frequency-domain audio representation that is shift-invariant for sustained notes. This representation is composed of the magnitude spectrum and the phase gradient, and can be later turned to audio via: 1. a phase integration algorithm and 2. the inverse STFT. The proposed design can be used with an efficient shift-invariant neural architecture and still yield stable reconstruction of sustained notes. Specifically, our contributions include:

- a formulation of the mel spectrogram inversion task, matching shift-invariant network and target, in order to improve the perceived stability of sustained notes
- a phase integration algorithm
- an evaluation metric measuring pitch stability for multiple notes

*Equal contribution

[1]Reconstruction examples https://machinelearning.apple.com/research/mel-spectrogram

2. BACKGROUND

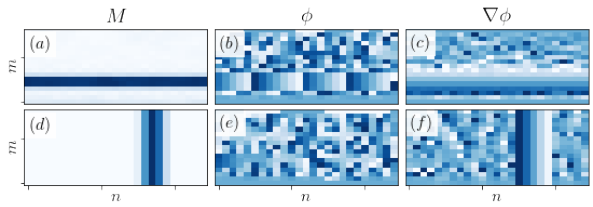


Figure 2: Magnitude M, phase ϕ and phase gradient $\nabla\phi$ patterns for a sinusoidal (top row) and an impulse (bottom row) signal. While the magnitude spectrum is easy to interpret visually, patterns in the phase spectrum are harder to decipher, but become evident in the phase gradient.

A discrete audio signal x can be analyzed in the time-frequency space using the STFT:

$$X[m,n] = \sum_i x[i+nR]w[i]e^{-j\omega_m i}, \tag{1}$$

where m and n are the integer frequency bin and time frame indices, R is the hop size between successive frames, w is a window function defined in the $[-N/2, N/2)$ interval, with N being the frame size and $\omega_m = 2\pi m/N$ the angular frequency. The STFT is a complex-valued matrix, as such it can be represented in the polar form:

$$X[m,n] = M[m,n]e^{-j\phi[m,n]}. \tag{2}$$

The magnitude component M highlights the energy of the signal at various locations in the time-frequency grid: it is easier to interpret and more widely used than the phase component ϕ. However, the phase spectrum is of primary perceptual importance to reconstruct the audio signal precisely, and while harder to interpret at first sight, it does contain patterns that can guide model design choices (Fig. 2).

In Sect. 2.1 we describe two patterns that form in the magnitude and phase spectrum corresponding to the occurrence of ideal sinusoidal and impulsive signal components.

2.1 Sinusoidal and impulsive components

2.1.1 Sinusoidal components

In the magnitude spectrum, sinusoidal components such as any single harmonic of a pitched instrument's sustained note, show up as horizontal lines (Fig. 2(a)). While the magnitude spectrum does not depend on the frame index n, and is therefore shift-invariant, the phase spectrum depends linearly on n (Fig. 2(b)), and the rate of change is given by the frequency of the sinusoidal component. Failing to reconstruct this linear relation between phase and time results in loss of horizontal phase coherence, perceived as unstable pitch, because errors are attributed to sudden changes of the frequency of the sinusoidal component.

2.1.2 Impulsive components

Impulsive components such as the attack of a percussion instrument show up as vertical lines in the magnitude spectrum (Fig. 2(d)). While the magnitude spectrum does not depend on the frequency index m, the phase spectrum depends linearly on m (Fig. 2(e)), and the rate of change depends on the offset between the location of the impulse and the frame center. Failing to reconstruct this linear relation between phase and frequency results in loss of vertical phase coherence, which is perceived as smeared transients, because the errors are attributed to the location of the impulse.

2.2 Phase gradient

The linear patterns that emerge in the phase for sinusoidal and impulsive components are better highlighted in the two components of the phase gradient $\nabla\phi = (\phi'_i, \phi'_m)$.

The partial derivative of phase along the time dimension ϕ'_i is called *instantaneous frequency*. For the bins that belong to sinusoidal components, ϕ'_i is constant and the phase can be propagated horizontally:

$$\phi[m, n+1] = \phi[m,n] + R\phi'_i[m,n]. \tag{3}$$

The partial derivative along the frequency dimension ϕ'_m is called *local group delay*. For the bins that belong to impulsive components, ϕ'_m is constant and the phase can be propagated vertically:

$$\phi[m+1, n] = \phi[m,n] + \phi'_m[m,n]. \tag{4}$$

In the time-frequency reassignment literature (see e.g. [7]), the phase gradient components are used to assign the energy of a spectral bin (m,n) to a nearby point of maximum contribution $(\dot{m}, \dot{n})$:

$$\begin{aligned} \dot{m}[m,n] &= m + \Delta m[m,n] \\ \dot{n}[m,n] &= n + \Delta n[m,n], \end{aligned} \tag{5}$$

where $\Delta n[m,n]$ (Fig. 2(f)) and $\Delta m[m,n]$ (Fig. 2(c)) represent time and frequency bin offsets and are derived from the phase gradient

$$\begin{aligned} \Delta m[m,n] &= \phi'_i[m,n]\tfrac{N}{2\pi} - m \\ \Delta n[m,n] &= -\phi'_m[m,n]\tfrac{N}{2\pi R}. \end{aligned} \tag{6}$$

In the following sections, we discuss how the phase gradient can be used for mel spectrogram inversion.

2.3 Mel spectrogram inversion

The log-amplitude mel spectrogram (simply mel spectrogram from now on) is a low-resolution time-frequency representation that is derived from the power spectrogram M^2, by first warping the frequency axis using the mel scale, then scaling the values to log-amplitude. Estimating the original audio signal x from the mel spectrogram requires recovering the information that has been lost in the direct computation, i.e. the phase information and