Advancing Transposition-Preserving Pitch Estimation in Audio Signals with Neural Network Architectures

Xincheng Zhang *

School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

* Corresponding Author Email: 1807050205@stu.hrbust.edu.cn

Abstract. This study addresses the challenge of transposition-preserving pitch estimation from audio signals, a critical task in music information retrieval that facilitates robust pitch recognition across varied musical transpositions. Accurate pitch estimation is foundational for applications like automated music transcription and music analysis, where maintaining the relational integrity of pitch shifts is essential. The objective of this research is to develop a neural network model capable of predicting pitch distributions from Constant-Q Transform (CQT) frames that are both original and pitch-shifted. The proposed model incorporates advanced neural architecture involving convolutional layers and transposition-preserving Toeplitz fully connected layers. Specifically, the model processes input through layer normalization and a series of 1D convolutions, integrating Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA) to enhance feature recognition capabilities. This is followed by a softmax layer for classification, ensuring the model's outputs reflect the transposed relationships evident in the inputs. Experiments conducted on a comprehensive audio dataset demonstrate that the integration of attention mechanisms significantly enhances model performance, with the Coordinate Attention module proving particularly effective in spatial feature recognition. The results highlight the capability to preserve pitch information across transformations and confirm its potential in real-world applications. This study not only advances the field of music information retrieval but also establishes a framework for future developments in pitch estimation technologies.

Keywords: Transposition-Preserving Pitch Estimation; Music Information Retrieval; Attention Mechanisms.

1. Introduction

Pitch estimation plays a critical role in various applications of audio processing, particularly in the domains of music information retrieval (MIR) and automatic music transcription (AMT) [1]. This is primarily due to pitch being a fundamental attribute that defines the tonal quality of sound and is crucial for comprehending and analyzing the structure and expressive elements of music. Advances in pitch estimation not only enhance the capabilities of MIR systems but also significantly contribute to fields such as speech recognition, linguistic analysis, and medical diagnostics where tonal analysis is essential [2]. Given its broad applicability and the intricate nature of audio signals, especially in polyphonic music [3], ongoing research in this domain is indispensable [4]. It serves to refine the accuracy of pitch detection algorithms and adapt these methodologies to various challenging audio environments.

Pitch estimation in AMT has undergone significant evolution due to advances in signal processing and machine learning techniques. Traditional methods such as autocorrelation and Fast Fourier Transform (FFT) initially dominated, providing foundational tools for pitch detection in monophonic and simple polyphonic scenarios [5]. However, these methods often struggle with the complex harmonic and temporal interactions found in real-world music recordings. Recent research has increasingly focused on machine learning approaches, particularly deep neural networks (DNNs), which can directly learn complex patterns from data [6]. Convolutional Neural Networks (CNNs) have been extensively utilized for their ability to extract hierarchical features from Spectro temporal representations of audio signals, thereby improving the accuracy of pitch estimation in both single-note and polyphonic contexts [7]. Studies have demonstrated the effectiveness of CNN-based

architectures in addressing challenges such as harmonic overlapping and transient noise suppression, crucial for precise transcription of musical content [8]. The emergence of deep learning has enabled newer models to significantly surpass the performance of traditional techniques. CNNs and autoencoders have demonstrated superior capability in handling complex tasks like multi-pitch estimation (MPE) in real-time applications, providing robustness against noise and achieving higher precision [9]. These advancements often build upon enhancements in dataset quality and diversity, along with innovations in model architectures like attention mechanisms and end-to-end learning strategies [10]. Furthermore, recent strides in self-supervised learning approaches have reduced reliance on large, annotated datasets, facilitating the training of pitch estimation models with limited labeled data [11]. Techniques such as contrastive predictive coding (CPC) and masked pitch prediction have been leveraged to enhance model generalization and robustness [12]. Moreover, recent advancements in this field have emphasized the importance of developing models that are equivariant to pitch transposition [13].

Self-supervised learning (SSL) offers a promising solution to the lack of labeled data in various domains. This research aims to advance pitch estimation by leveraging SSL and enhancing the Pitch Estimation with the Self-supervised Transposition-equivariant Objective (PESTO) model, comparing its performance with baseline methods. The PESTO model refines its predictions by comparing two versions of the same sound that have been randomly but consistently transposed in pitch. This approach allows the model to rely on internal consistency checks rather than extensive labeled data. The network processes original and pitch-shifted Constant-Q Transform (CQT) frames, applying a combination of layer normalization, 1D convolutional layers with skip connections, and transposition-preserving Toeplitz layers.

Extensive experiments evaluate the proposed method, integrating advanced neural network techniques such as Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA) modules [14]. Unlike previous approaches, the class-based equivariance loss function prevents model collapse—a common challenge in SSL frameworks—without needing a separate decoder. Additionally, the model architecture employs Toeplitz fully connected layers, ensuring it remains lightweight and transposition-equivariant. The significance of this research lies in its potential to advance state-of-the-art pitch estimation. By utilizing attention mechanisms and improving the PESTO algorithm, this approach can overcome the limitations of other methods, providing more accurate and robust pitch estimation. These advancements hold significant practical value for real-time music analysis applications, enhancing interactive music systems and automated tools for digital music production.

2. Methodology

2.1. Dataset Description and Preprocessing

The MIR-1K dataset comprises 1,000 song clips with pitch contours [15], ideal for melody extraction and pitch analysis. This dataset consisted of recordings where all singing voices and musical accompaniments were recorded separately. Each song clip was manually labeled with pitch values, unvoiced sounds, vocal/non-vocal segments, lyrics, and the speech recording of the lyrics. The network processes CQT frames computed by nnAudio's CQT module, the hop duration is 10ms. CQT is preferred in music and audio processing as it provides a time-frequency representation where the frequency bins are geometrically spaced, the pitch of the CQT representation randomly been shifted within a specified range, then complex CQT values been converted to log magnitude, the frequency range from 27.5Hz to 16kHz. The logarithm of the magnitude (in dB) is used as it closely resembles human perception of sound intensity. The data augmentation involves two parts, including adding random noise to a batch of audio samples as well as randomly changing the gain of audio samples. The noise level is regulated by a signal-to-noise ratio (SNR), which is randomly sampled for each audio sample in the batch. White noise with a standard deviation between 0.1 and 2 and gain with a random value uniformly selected between −6 and 3 dB, are employed.

2.2. The Model Transposition-Preserving Operations

The model begins with a CQT frame of the audio signal. The original CQT frame is transposed by a specified number of semitones, x (k), to create a variant of the input that mimics a pitch shift. Both the original and pitch-shifted CQT frames are further augmented (denoted as $\tilde{x}$ and $\tilde{x}$ (k)), as shown in the Fig. 1. This is followed by a deep residual network with multiple 1D convolutional layers, each convolutional output is layer normalized, non-linear activations as Leaky- Rectified Linear Unit (ReLU) with a slope of 0.3 and dropout of 0.2 are employed to introduce non-linearity into the model and prevent overfitting. Skip-connections are integrated to help the network learn identity functions across layers. The final part of the network architecture incorporates a Toeplitz fully connected (FC) layer. Toeplitz matrices are used because of their transposition-preserving properties—important for tasks where recognizing pitch shifts in audio is crucial. The model proposes as depicted in Fig. 2.

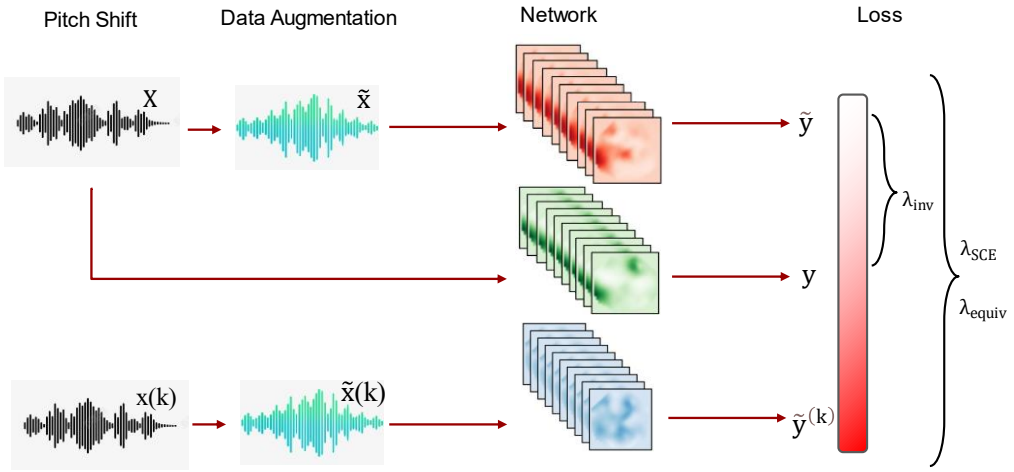

Fig 1. Overview of the method.

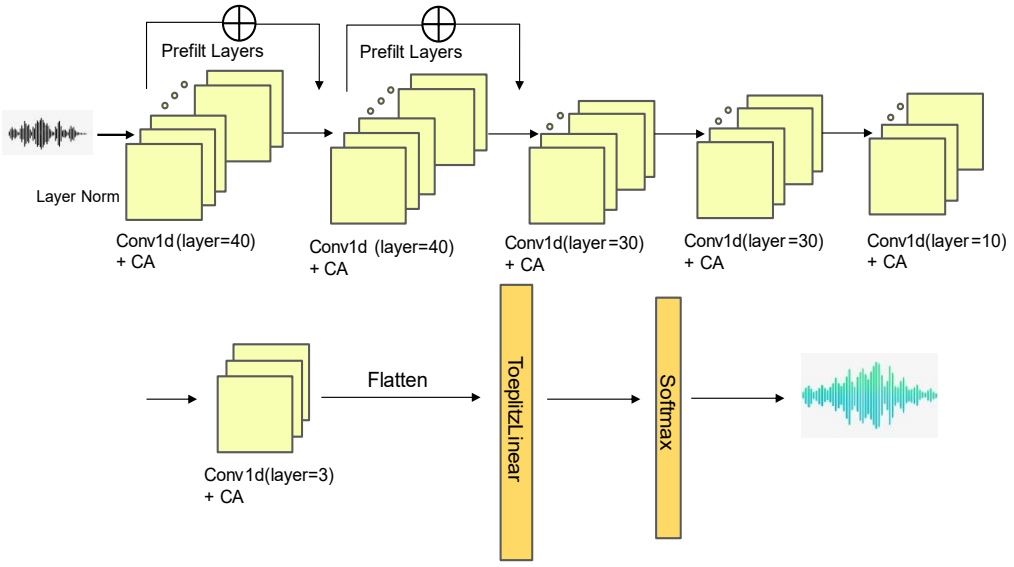

Fig 2. The pipeline of the model.

2.2.1. Class-Based Transposition-Equivariant Objective

By framing the transposition as a class-based task, the model is encouraged to differentiate between different classes of pitch shifts. Each class corresponds to a different transposition interval. The class-based objective transforms the task of understanding pitch shifts into a classification problem within a SSL framework. This prevents the output from collapsing to a single value or a few values, as doing so would result in poor performance on this classification task. Convolutions inherently preserve transpositions because the same kernel applied over shifted inputs will produce correspondingly shifted outputs.

Moreover, the objective focuses on the preservation of pitch information in the embeddings produced by the encoder. Unlike standard FC layers, a Toeplitz matrix ensures that each output neuron's connection pattern is a shifted version of the others, perfectly aligning with the requirement to preserve pitch transformations in the output. The matrix's constant diagonal values ensure consistent treatment of similar shifts across all inputs, making the network sensitive to pitch but invariant to where exactly these pitches occur in the input. This focus mandates that the encoder retains a high degree of sensitivity to pitch changes in the input, ensuring that pitch information is not lost but rather emphasized in the learning process. This approach also enhances the model's robustness, minimizes the risk of the model overfitting to noise or other irrelevant details in the data by directing the model's learning focus towards pitch.

2.2.2. PESTO with CA module

A critical innovation is the Toeplitz FC layers. Unlike standard FC layers, Toeplitz FC layer weights form a Toeplitz matrix, making each descending diagonal constant. This equates the FC layer operation to a 1D convolution, preserving transposition equivalence. The convolution layers' output is flattened before the Toeplitz FC layer, linking feature extraction and classification seamlessly. The output from the Toeplitz FC layer passes through a SoftMax layer, converting it into a probability distribution over possible pitch class. The SoftMax function highlights the most probable pitch class, ensuring output probabilities sum to one. The model's architecture is significant due to the use of layer norm and residual connections, managing the vanishing gradient problem and enabling deeper networks. The Toeplitz matrices in FC layers maintain transposition equivalence, aligning with music's physical properties.

The CA module is added right after each convolutional layer (including prefiltering layers) to refine features through channel and spatial attentions sequentially. In the revised CA module for 1D data, the architecture is refactored to streamline feature processing and enhance model efficiency. The original multidimensional CA module calculates attention as:

$$Xout = X \odot \sigma(Conv(W, [AvgPoolh(X), AvgPoolw(X)])) \tag{1}$$

For 1D applications, it simplifies to:

$$Xout = x \odot \sigma(Conv(W, AvgPool(x))) \tag{2}$$

This involves a pooling layer that compresses each feature channel into a single scalar, effectively summarizing the channel's content. Subsequently, a reduction layer diminishes the channel dimensionality, reducing both the parameter count and computational complexity. This is followed by the activation function Mish, which introduces non-linearity after normalization. An expansion layer then restores the channel dimensions from the reduced representation, culminating in the application of a sigmoid function to generate attention weights. These weights are used to scale the input channels proportionally to their deemed importance. This two-step process of reduction and expansion ensures that the module maintains a compact channel-wise descriptor while accurately modulating the input by emphasizing or de-emphasizing channels based on learned attention, thereby optimizing the channel-specific response to features in 1D data environments.

2.2.3. Loss Function

Equivariance Loss is a loss function that penalizes the model when the ratio of outputs between the original and pitch-shifted inputs does not match the expected ratio derived from the pitch shift amount. The loss measures the invariance to transformation between two probability distributions. This can be expressed by q (k*w) = k * q (w), where q (w) is the original representation. Using a deterministic linear projection—power series transformation to reduce high-dimensional data to a scalar that represents the probability distribution of pitches. Therefore, the loss function is expressed as:

$$\lambda_{equiv}(y, y^{(k)}, k) = h_\tau(\frac{\phi(y^{(k)})}{\phi(y)} - \alpha^k) \tag{3}$$

The h_τ represents the Huber loss function:

$$h_\tau(x) = \begin{cases} \frac{x^2}{2} & if|x| \leq \tau \\ \frac{\tau^2}{2} + \tau(|x| - \tau) & otherwise \end{cases} \quad (4)$$

The Softmax Cross-Entropy Loss (λ_{SCE}) between the two probability distributions are calculated with the following formula, where k stands for the k transposed pitches. Typically used in classification tasks to ensure the predicted probability distribution matches the true distribution.

$$\lambda_{SCE}(y, y^{(k)}, k) = \sum_{i=0}^{d-1} y_i \, log(\, y_{i+k}^{(k)}) \quad (5)$$

Invariance Loss (L_{inv}) is used to ensure that timbral characteristics remain unaffected by the pitch-shifting, where the model is expected to predict the same pitch distribution given a slightly modified version of the sound.

$$\lambda_{inv}(y, \tilde{y}) = CrossEntropyLoss(y, \tilde{y}) \quad (6)$$

Different losses are combined to allow the model to learn both absolute pitch and relative shifts. The loss function used can be expressed as:

$$L = \lambda_{inv} L_{inv}(y, \tilde{y}) + \lambda_{equiv} L_{equiv}(\tilde{y}, \tilde{y}^{(k)}, k) + \lambda_{SCE} L_{SCE}(\tilde{y}, \tilde{y}^{(k)}, k) \quad (7)$$

3. Result and Discussion

3.1. Loss-Weight

Increasing L_{inv} would make the network focus more on being robust to non-pitch-related variations in the input. Increasing λ_{equiv} or λ_{SCE} would make the network focus more on understanding and preserving the pitch-related information across transformations and pitch shifts, as depicted in Fig. 3. The weight of λ_{inv} is 20%, and the weight of λ_{equiv} is 40%, as well as the weight of λ_{SCE}.

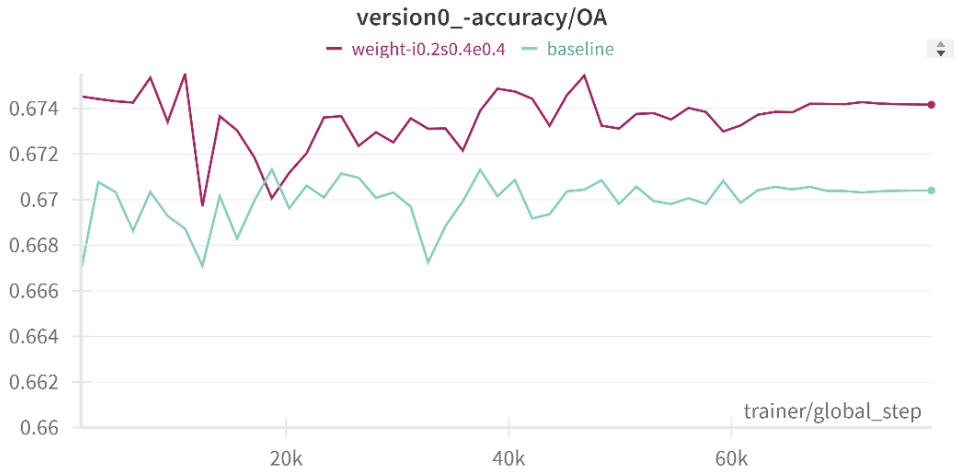

Fig 3. The performance of changing the loss-weight.

3.2. Contribution of various Attention Modules

The impact of integrating SE, CBAM, and CA modules into the model's encoding layers is examined, analyzing their performance variations and implications.

The OA accuracy results depicted in Fig. 4 illustrate the comparative performance of the SE, CBAM, and CA modules against the baseline. The CA module consistently outperforms the baseline, highlighting its effectiveness in capturing spatial dependencies crucial for melody extraction tasks. The CBAM module shows higher initial variance but stabilizes over training iterations to perform

better than the baseline, demonstrating its efficient use of both channel and spatial features. Interestingly, the combination of SE+CA mirrors the baseline performance, suggesting that the potential synergy expected from this combination does not translate into tangible benefits over using CA alone in this specific setup. As the RCA accuracy results shown in Fig. 3, the CA module again leads with higher peaks, indicating superior handling of octave errors. This outcome underscores CA's precise spatial awareness, which is pivotal for chroma accuracy in complex musical pieces. CBAM's performance trends closely with CA, benefiting from its dual attention mechanisms that effectively capture essential frequency nuances. Conversely, the SE+CA module underperforms compared to both CA and CBAM, which might be attributable to redundancy in attention mechanisms. The RPA accuracy results reveals that the CA module frequently surpasses other methods, emphasizing its strengths in addressing pitch-specific accuracy within stringent error tolerances. The CBAM module showcases comparable performance, reinforcing its capability to balance spatial and channel dependencies effectively.

3.3. Discussion of the mechanisms

As the params shown in Table 1, the primary purpose of using a kernel size = 1 in CA modules is to perform transformations across the channels of the input tensor without altering the spatial (or sequential, in the case of 1D convolutions) dimensions.

Fig. 4 Performance with different attention modules

Table 1. The params of the modules with accuracy results.

	Trainable params	Kernel size	Reduction-RT	OA accuracy	RCA accuracy	RPA
CA(full)	31.3k	1	4	67.43%	96.25%	95.93%
SE+CA	29.4k	1	16+6	67.28%	96%	95.69%
CBAM	31.5k	9	4	67.54%	96.03%	95.67%
Base	28.3k	15	NaN	67.27%	96.12%	95.68%

Also, using kernel size = 1 is computationally efficient as it simplifies the convolution operation to a mere element-wise multiplication and addition across channels. Larger kernel size (e.g., 9x1 in CBAM) can capture more spatial context and interactions between more distant elements within the input data. This can be particularly beneficial for audio signals where contextual information might be relevant over a larger temporal span.

The examination of SE, CBAM, and CA modules across various accuracy metrics clearly demonstrates the superiority of CA in spatial feature recognition, crucial for effective pitch estimation. CBAM's balanced approach between channel and spatial attentions proves beneficial yet does not consistently surpass the more focused CA module. SE, particularly when combined with CA, shows limited impact, suggesting that its channel recalibration does not complement CA's spatial attention in a way that enhances overall model performance. This analysis not only underscores the importance of selecting appropriate attention mechanisms based on task-specific needs but also affirms the significant improvements these modules bring to pitch estimation tasks in real-world scenarios.

4. Conclusion

This study addresses the challenge of pitch estimation in audio processing by introducing an innovative neural network architecture designed to maintain transposition equivalence, which is essential for accurately predicting the pitch distribution of transposed CQT frames. The proposed method employs a nuanced approach, incorporating CBAM, SE, and CA within the convolutional layers of the model. These modules are intricately integrated to enhance the ability to recognize and process spatial and channel-wise features effectively, with each layer meticulously structured to support the primary goal of transposition-preserving pitch prediction. Extensive experiments were conducted to evaluate the effectiveness of the proposed method, particularly focusing on the integration of attention mechanisms and their impact on the model's performance. The experimental results highlight the superiority of the CA module in spatial feature recognition, which is crucial for precise pitch estimation. The CA module's focused approach to handling spatial details consistently enhanced pitch accuracy, outperforming other configurations. Looking ahead, the study will focus on improving the robustness and generalization capabilities of the model under varied and noisy audio environments. Future research will aim to refine the model's ability to handle non-pitch-related variations such as timbral and dynamic range changes, ensuring the applicability in more diverse real-world scenarios. This will involve adjusting the loss components to fine-tune the balance between pitch preservation and insensitivity to non-essential variations, driving further advancements in audio processing technology.

References

[1] Kim J.W. Automatic Music Transcription in the Deep Learning Era: Perspectives on Generative Neural Networks. New York University, 2020.

[2] Hess W. Pitch determination of speech signals: algorithms and devices. Springer Science & Business Media, 2012.

[3] Duan Z. Temperley D. Note-level Music Transcription by Maximum Likelihood Sampling. ISMIR, 2014: 181-186.

[4] Gowrishankar B.S. Bhajantri N.U. An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques. International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). IEEE, 2016: 140-152.

[5] Matsunaga T. Saito H. Multi-Layer Combined Frequency and Periodicity Representations for Multi-Pitch Estimation of Multi-Instrument Music. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[6] Sigtia S. et al. An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(5): 927-939.

[7] Elghamrawy S.M. Ibrahim S.E. Audio signal processing and musical instrument detection using deep learning techniques. International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC). IEEE, 2021: 146-149.

[8] Elowsson A. Friberg A. Modeling music modality with a key-class invariant pitch chroma CNN. 2019, arXiv preprint: 1906.07145.

[9] Morrison M. Hsieh C. Pruyne N. et al. Cross-domain neural pitch and periodicity estimation. 2023, arXiv preprint: 2301.12258.

[10] Wu Y.T. Chen B. Su L. Multi-instrument automatic music transcription with self-attention-based instance segmentation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2796-2809.

[11] Gfeller B. Frank C. Roblek D. et al. SPICE: Self-supervised pitch estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1118-1128.

[12] Wang C. Li Z. Tang B. et al. Towards high-fidelity singing voice conversion with acoustic reference and contrastive predictive coding. 2021, arXiv preprint: 2110.04754.

[13] Riou A. Lattner S. Hadjeres G. et al. Pesto: Pitch estimation with self-supervised transposition-equivariant objective. International Society for Music Information Retrieval Conference (ISMIR 2023). 2023.

[14] Hou Q. Zhou D. Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[15] Hsu C.L. Jang J.S.R. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. IEEE transactions on audio, speech, and language processing, 2009, 18(2): 310-319.