# Switching Machine Improvisation Models by Latent Transfer Entropy Criteria

Shlomo Dubnov, Vignesh Gokul, Gérard Assayag

physical sciences forum

MDPI

*Proceeding Paper*

# Switching Machine Improvisation Models by Latent Transfer Entropy Criteria †

**Shlomo Dubnov [1,*], Vignesh Gokul [2] and Gerard Assayag [3]**

[1] Department of Music, UC San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA
[2] Department of Computer Science, UC San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA
[3] Institut de Recherche et de Coordination Acoustique/Musique, 1 Place Igor Stravinsky, 75004 Paris, France
* Correspondence: sdubnov@ucsd.edu
† Presented at the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris, France, 18–22 July 2022.

**Abstract:** Music improvisation is the ability of musical generative systems to interact with either another music agent or a human improviser. This is a challenging task, as it is not trivial to define a quantitative measure that evaluates the creativity of the musical agent. It is also not feasible to create huge paired corpora of agents interacting with each other to train a critic system. In this paper we consider the problem of controlling machine improvisation by switching between several pre-trained models by finding the best match to an external control signal. We introduce a measure SymTE that searches for the best transfer entropy between representations of the generated and control signals over multiple generative models.

**Keywords:** generative models; transfer entropy; granger causality; musical information dynamics



**Citation:** Dubnov, S.; Gokul, V.; Assayag, G. Switching Machine Improvisation Models by Latent Transfer Entropy Criteria. *Phys. Sci. Forum* **2022**, *5*, 49. https://doi.org/10.3390/psf2022005049

Academic Editors: Frédéric Barbaresco, Ali Mohammad-Djafari, Frank Nielsen and Martino Trassinelli

Published: 8 February 2023



## 1. Introduction

Learning generative models of complex temporal data is a formidable problem in Machine Learning. In domains such as music, speech or video, deep latent-variable models manage today to generate realistic outputs by sampling from predictive models over a structured latent semantic space. The problem is often further complicated by the need to sample from non-stationary data where the latent features and its statistics change over time. Such situations often occur in music and audio generation, since musical structure and the type or characteristics of musical sounds change during the musical piece. Moreover, in interactive systems the outputs need to be altered so as to fit user specifications, or to match another signal that comes from the environment, which provides the context or constraint for the type of desired outcome produced by the generative system at every instance. In such cases generation by conditional sampling might be impossible due to lack of labeled training data and the need to retrain the models for each case.

We call this problem Improvisation Modeling, since it is often encountered in musical interaction with artificial musical agents that need to balance their own artificial "creativity" with responsiveness to the overall musical context in order to create a meaningful interaction with other musicians. The ability of the artificial musical agent to make decisions and switch its responses by listening to a human improviser is important for establishing the conditions for man-machine co-creation. We consider this as a problem of controlling machine improvisation by switching between several pre-trained models by finding the best match to an external context signal. Since the match can be partially found in different generative domains, we search for best transfer entropy between reduced representations of the generated and context signals across multiple models. The added step of matching in the reduced latent space is one of the innovations of the proposed method, also motivated by theories of cognition that suggest mental representation as lossy data encoding.

In order to allow quantitative analysis of what is happening in the "musical mind", we base our work on an information theoretic music analysis method of Music Information Dynamics (MID). MID performs structural analysis of music by considering the predictive aspects of music data, quantified by the amount of information passing from past to present in a sound recording or symbolic musical score. We extend the MID idea to include the relation between the generated and context signals and their latent representations, amounting to a total of five factors: the signal past X with its latent encoding Z, the signal present sample Y, a context signal C and its encoding into latent features T. Assuming Markov chain relations between Z-X-Y, we are looking for the smallest latent representation Z that predicts the present Y, while at the same time having maximal mutual information to the latent features T of the constraint signal. For each model we compute transfer entropy between the generated and context latent variables Z and T, respectively, and the present sample Y. It should be noted that our notion of Transfer Entropy is different from the standard definition of directed information between two random variables, since transfer entropy is estimated in the latent space of the generative model and the context signal.

We propose the use of a new metric called Symmetric Transfer Entropy (SymTE) to switch between multiple pre-trained generative models. This means that given any audio context signal, we can use SymTE to effectively switch between multiple outputs of generative models. In the paper we will present the theory and some experimental results of switching pre-trained models according to second musical improvisation input. An important aspect of our model is eliminating the need to re-train the temporal model at each compression rate of Z since estimation of I(Y,Z) is not needed for model selection. Our assumption is that we have several pre-trained generative models(or random generators), each providing one of multiple options for improvised generation. The best model is chosen according to criteria of highest latent transfer entropy by search for the optimal reduced rate for every model, balancing between the quality of signal prediction (predicting Y from full rate Z) and matching between the past latent representation of Z and the latent representation T of the context signal for that model.

### *1.1. Causal Information*

The problem of inferring causal interactions from data was formalized in terms of linear autoregression by Granger [1]. The information-theoretic notion of transfer entropy was formulated by Schreiber [2] not in terms of prediction, like in the Granger case, but in terms of reduction of uncertainty, where transfer entropy from Y to X is the degree to which Y reduced the residual uncertaintly about the future of X after the past of X was already taken into consideration. It can be shown that Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables [3] Causal entropy $\Sigma_{t=1}^{T} H(Y_t|X_{1:t}, Y_{1:t-1})$ measures the uncertainty present in the conditioned distribution of the $Y$ variable sequence given the preceding partial $X$ variable sequence [4].

It can be interpreted as the expected number of bits needed to encode the sequence $Y_{1:t}$ given the sequentially revealed previous $Y$ variables and side information, $X_{1:t}$. Causal information (also known as the directed information) is a measure of the shared information between sequences of variables when the variables are revealed sequentially $\Sigma_{t=1}^{T} I(Y_t; X_{1:t}|Y_{1:t-1})$ [5]. Transfer Entropy is closely related to Causal information, except for considering the influence on $Y$ from past of $X$ only, not including the present instance t. Moreover, in some instances the past of $X$ is considered for shorter past, or even just a single previous sample.

Understanding causality is important for man-machine co-creativity, especially in improvisational settings, since creating a meaningful interaction also requires answering the question of how does the human mind go beyond the data to create an experience [6]. In a way, the current work goes beyond the predictive brain hypothesis [7] to address issues of average predictability and of reduced representation of sensations as "hidden causes" or "distal causes" that maximize the communication between human and a machine in improvisational setting.

### *1.2. Estimating Transfer Entropy*

Several tools and methods have been proposed to estimate transfer entropy. Refs. [8,9] use entropy estimates based on k-nearest neighbours instead of conventional methods such as binnings to estimate mutual information. This can be extended to estimating transfer entropy, as transfer entropy can also be expressed as conditional mutual information. Similarly, methods based on Bayesian estimators [10] and Maximum Likelihood Estimation [11,12] proposed a method to estimate transfer entropy based on Copula Entropy.

Methods using neural networks have also been proposed to estimate mutual information. Mutual Information Neural Estimator (MINE) [13] estimate mutual information by performing gradient descent on neural networks. Intrinsic Transfer Entropy Neural Estimator (ITENE) [14] proposes a two-sample neural network classifiers to estimate transfer entropy. Their method is based on variational bound on KL-divergence and pathwise estimator of Monte Carlo gradients.

Several toolboxes and plugins such as Java Information Dynamics Toolkit (JIDT) [15] provide implementations of the above mentioned methods. However, most of them have not been tested on complex high-dimensional data such as music. To our best knowledge, we are the first to propose a transfer entropy estimation method on complex data such as music and demonstrate results on tasks such as music generation.

## 2. Methodology

The main objective of our work is to calculate a metric based on transfer entropy to switch between outputs of different generative processes ( say $X_1, X_2, ...X_N$), so that the output is semantically meaningful to a context signal ($C$). For a given $X_i$, we denote the past by $\bar{X}_i$ and similarly we denote the past of $C$ as $\bar{C}$.

Transfer Entropy between two sequences is the amount of information passing from the past of one sequence to another, when the dependencies of the past of the other sequence (the sequence own dynamcis) have been already taken into account. In the case of $C$ and model's $i$ data $X_i$ we have $TE_{C\to X_i} = I(X;\bar{C}|\bar{X})$ Similarly $TE_{X\to C} = I(C;\bar{X}|\bar{C})$. Writing mutual information in terms of entropy

$$I(C;\bar{X}_i) = H(C) - H(C|\bar{X}_i)$$
$$I(C;\bar{X}_i|\bar{C}) = H(C|\bar{C}) - H(C|\bar{X}_i,\bar{C})$$

Adding and subtracting H(C):

$$I(C;\bar{X}_i|\bar{C}) = H(C|\bar{C}) - H(C|\bar{X}_i,\bar{C}) - H(C) + H(C) = I(C;\bar{X}_i,\bar{C}) - I(C;\bar{C}) \quad (1)$$

Also:

$$I(X_i;C|\bar{X}_i) = I(X_i;\bar{C},\bar{X}_i) - I(X_i;\bar{X}_i) \quad (2)$$

We consider a sum of (1) and (2), let's call it symmetrical transfer entropy(SymTE):

$$SymTE = I(C;\bar{X}_i|\bar{C}) + I(X_i;\bar{C}|\bar{X}_i) \quad (3)$$

As shown in the Appendix A, one can derive the following equivalent expression

$$SymTE = I((C,X_i);\overline{(C,X_i)}) - I(C;X_i|\overline{(C,X_i)}) + I(C,X_i) - I(X_i,\bar{X}_i) - I(C,\bar{C})\ , \quad (4)$$

where we used a notation for past of the joint pair $(\bar{C},\bar{X}_i) = \overline{(C,X_i)}$.

The measure of mutual information between the present and the past of a signal, known as information rate (IR), will be explained in the next section. IR is commonly used in analysis of Music Information Dynamics (MID) that captures the amount of average surprisal in music signals when the next sound event is anticipated from its past. If we

assume that the generation of $X_i$ is independent of $C$ given their joint past $\overline{(C, X_i)}$, then $I(C; X_i|\overline{(C, X_i)}) = 0$, resulting in

$$SymTE \approx I((C, X_i); \overline{(C, X_i)}) - I(X_i, \bar{X}_i) - I(C, \bar{C}) + I(C, X_i)\,, \tag{5}$$

which is a sum of IR of the joint pair $(C, X_i)$ and the mutual information between $C$ and $X_i$ regardless of time, minus IR of the separate stream. In other words, the Symmetrical TE is a measure of surprisal present in the joint stream minus the surprisal of each of its component, plus the mutual information (lack of independence) between the individual components. In a way this captures the difference between predictive surprisal when listening to a compound stream versus surprisal when listening separately, with added component of mutual information between the voices regardless of time.

This process is schematically represented in Figure 1 as a combination of Information Rate and Mutual Information estimates for two musical melodies

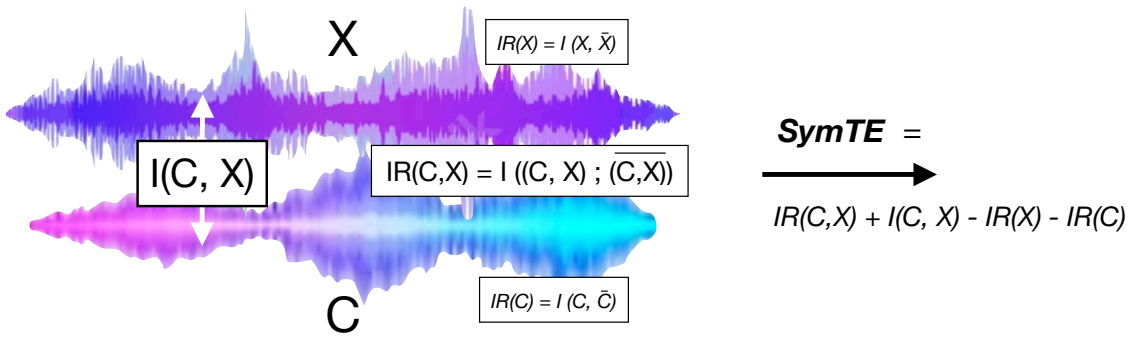


**Figure 1.** Estimate of *SymTE* as a combination of Information Rate *IR* and Mutual Information *I* estimates from a generated $X$ and control signal $C$.

### *2.1. Predictive Surprisal Using VMO*

The essential step in estimating the predictive surprisal is building a model called Variable Markov Oracle (VMO). Based on Factor Oracle (FO) string matching algorithm VMO was decveloped to allow generative improvisation for real-valued scalar or vector data, such as sequences of audio feature vectors, or data vectors extracted from human poses during dance movements. VMO uses suffix data structure for query-guided audio content generation [16] and multimedia query-matching [17,18]. VMO operates on multivariate time serie data, *VMO* symbolizing a signal $X$ sampled at time $t$, $X = x_1, x_2, \ldots, x_t, \ldots, x_T$, into a sequence $S = s_1, s_2, \ldots, s_t, \ldots, s_T$, having $T$ states and observation frame $x_t$ labeled by $s_t$. The labels are formed by following suffix links along the states in an oracle structure, whose value is one of the symbols in a finite sized alphabet $\Sigma$.

Predictive surprisal is estimated by constructing an FO automata for different threshold when search for suffix links. At each threshold value, a different oracle graph is created, and for each such oracle, a compression method of Compror (Compression Oracle) [19] algorithm $C$ is used as an approximation to predictive information $I(X, Y) = H(Y) - H(Y|X) \approx C(Y) - C(Y|X)$. Here the entropy $H$ is approximated by a compression algorithm $C$, and $C(Y) = log_2(|S|)$ is taken as the number of encoding bits for individual symbols over alphabet $S$, and $C(Y|X)$ is the number of bits in a block-wise encoding that recursively points to repeated sub-sequences [17].

As mentioned in the introduction, one of the advantages of using VMO for mutual information estimation is that it allows instantaneous time-varying estimates of IR based on the local information gain of encoding a signal based on linking it to its similar past. This differs from other methods of mutual information estimation like MINE that averages over the whole signal.

### *2.2. Border Cases*

If $C = X$, and since $H(X, X) = H(X)$, we get $I((X, X); \overline{(X, X)}) = IR(X)$ and $SymTe = I(X, X) - IR(X) = H(X) - H(X) + H(X|\bar{X}) = H(X|\bar{X})$, which is the conditional entropy of $X$ given its past. So TE of a pair of identical streams is its entropy rate.

If $C$ and $X$ are independent, $SymTE = 0$. This is based on the ideal case of IR estimator of the joint sequence $I((C, X_i); \overline{(C, X_i)})$ being able to reveal the IR of the individual sequences, and additionally capture any new emerging structure resulting from their joint occurrence.

In theory, if $C$ and $X$ are independent, $H(C, X) = H(C) + H(X)$, and $H(C, X|\overline{(C, X)}) = H(C|\bar{C}) + H(X|\bar{X})$, so $I((C, X_i); \overline{(C, X_i)}) = I(X_i, \bar{X}_i) + I(C, \bar{C})$. Thus, a combination of two streams may add additional information, but in practice it could be that VMO will not be able to find sufficient motifs or additional temporal structure when a mix is done. In such a case it can be that $SymTE$ estimate will become negative.

## 3. Representation Using VQ-VAE

Computing Information Rate and Mutual information for raw audio signals is an extremely challenging and computationally expensive task. We need some form of dimensionality reduction that preserves the semantic meaning of audio (style, musical rules, composer attributes etc) in the latent space. Then, we can estimate IR and MI in the lower dimensional space, quite easily. For our framework, we use a pre-trained Jukebox's Vector Quantized-Variational Autoencoder (VQ-VAE) [20,21] to encode raw audio files to low-dimensional vectors. VQ-VAE is a type of variational autoencoder that encodes data into a discrete latent space. These discrete codes correspond to continuous vectors in a codebook. Using this, we transform our data into 8192 64-dimensional latent vectors.

## 4. Switching between Generative Models

In this section, we explain the overall workflow of our method Figure 2. The main objective of our method is to switch between N different generative models to match a given query C. Given training data points (musical sequences), $D_1, D_2, ..., D_N$, we compute latent representations/embeddings of each data point to get $E_1, E_2, ..., E_N$. For our method, we use the embeddings of a pretrained VQ-VAE encoder from Jukebox. We construct each generative model $i$ as follows: (1) First we convert the query musical signal to the same latent space using Jukebox's VQ-VAE. (2) We create a $VMO_i$ for datapoint $i$ (in our case, we assume each datapoint represents a different composer). (3) Finally, we get the output of generative model $i$ by querying $VMO_i$ with embeddings of C to get $X_i$, algorithm provided in [16].

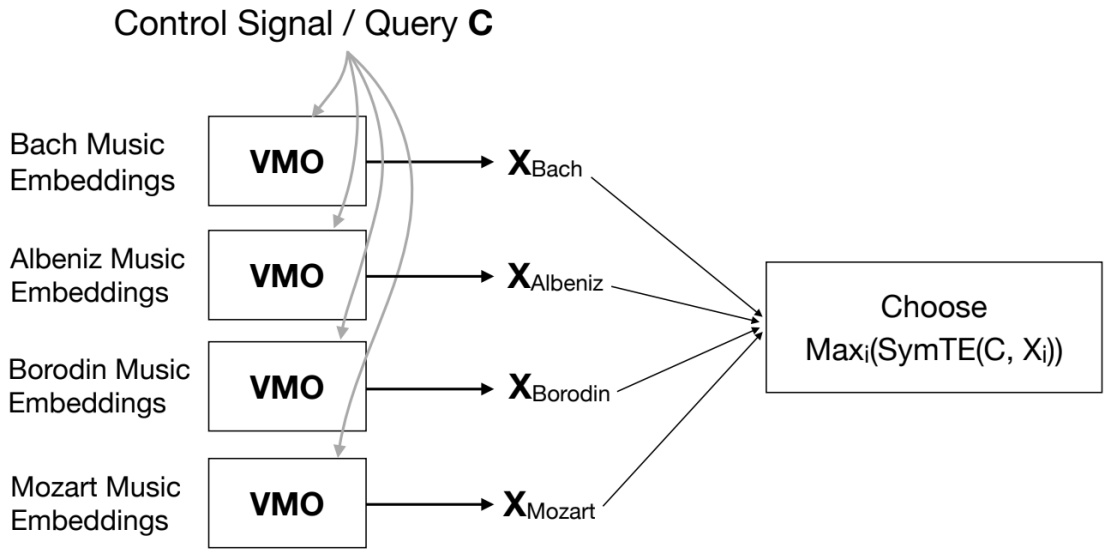


**Figure 2.** Our Methodology. Given music embeddings, we construct a VMO for each composer. For a control signal C, we query each composer's VMO to get $X_i$. We switch to the output $X_i$ for a control singal C, if $SymTE(C, X_i)$ is maximum for $i$.

In order to choose the best output for a given query C, we calculate $SymTE(X_i, C)$ for all $i \in 1, ..., N$. To calculate $SymTE(C, X_i)$, we need to calculate the individual terms of Equation 5, $I(C, \bar{C})$, $I(X, \bar{X})$, $I(C, X_i)$ and $I((C, X_i); \overline{(C, X_i)})$. To calculate $I(C, \bar{C})$, $I(X, \bar{X})$, we use VMO algorithm to create an oracle based on $X_i$ and another oracle for C to retrieve the information rate. To calculate $I(C, X_i)$, we use MINE to calculate the mutual information between C and $X_i$. To calculate $I((C, X_i); \overline{(C, X_i)})$, we propose two methods, we combine both $X_i$ and $C$ to create a mixture, based on two methods concatenation and addition of the respective latent vectors. Then, we create an oracle for the combined $(C, X_i)$ to calculate the IR.

## 5. Experiments and Results

We show the advantage of our method compared to other baselines by running simulations on the Labrosa APT dataset [22]. We construct a dataset with audio wav files of 4 different composers (Bach, Albeniz, Borodin, Mozart). We convert each audio file to the corresponding embeddings from a pre-trained Jukebox VQVAE [20,21]. For $X_i$, we create a VMO for each composer, that can synthesize a sequence of embeddings for a given query/context signal $C$. For our simulations, we construct $C$ as a segment of music (not included in the VMO construction) from any of the composers. Ideally, our SymTE measure should be high for the $X_i$ (VMO) of the same composer $C$.

We evaluate the effectiveness of SymTE by measuring the accuracy and F1 score. We conducted 20 trials for all the experiments. Each trial consisted of 20 query/context signals, randomly sampled from either of the 4 composers. For our baselines, we choose a random baseline and another baseline based on the euclidean distance of the embeddings, i.e., choose the composer i's output, for which the euclidean distance between the embeddings of C and embeddings of $X_i$ is the minimum. Table 1 shows the results of our methods and the baselines. We compare both averaging(avg) and concatenating (concat) the sequences in our experiments.We observe that our concat method achieves the best accuracy and F1-Score compared to all our baselines.

**Table 1.** Comparison of Accuracy and F1-Score of our methods and baselines.

| Method | Accuracy | F1-Score |
|---|---|---|
| Random | 0.22 | 0.17 |
| Distance-based | 0.27 | 0.17 |
| Our Method (concat) | **0.44** | **0.28** |
| Our Method (avg) | 0.36 | 0.21 |

## 6. Discussion and Future Work

The methods presented in the paper use sequence of latent vectors coming from pre-trained neural models of audio. We use VQ-VAE's embeddings and not the quantized codes, so there is only one quantization happening in this work, which is the VMO's. The reason why we chose VQ-VAE over other models is that we need strong pre-trained models and the best one currently is considered to be jukebox's VQ-VAE. Other neural models can be explored as well, as the representation is important for estimation of TE. Our query signals are 256 dimensional (≈0.71 s). Our method should work for longer queries, but the main bottleneck is the complexity of the generative model. We plan to extend this work with more elaborate results with a bigger data set and query size. We also plan to test the framework in terms of computational time, so as to enable real-time switching for music improvisation.

**Author Contributions:** Conceptualization, S.D., G.A. and V.G.; Software: S.D. and V.G.; Writing: S.D., V.G. and G.A.; Funding Acquisition: S.D. and G.A.; Project Administration: S.D. and G.A.; Supervision: S.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has received funding from the European Research Council (ERC REACH project) under the European Union's Horizon 2020 research and innovation programme (Grant agreement #883313).

**Institutional Review Board Statement:** Not Applicable.

**Informed Consent Statement:** Not Applicable.

**Data Availability Statement:** Data available in a publicly accessible repository that does not issue DOIs Publicly available datasets were analyzed in this study. This data can be found here: http://labrosa.ee.columbia.edu/projects/piano/.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

We provide a proof that Symmetric Transfer Entropy (SymTE), between two sequences $X$ and $C$, defined as a sum of individual Transfer Entropies given by

$$SymTE = I(C; \bar{X}|\bar{C}) + I(X_i; \bar{C}|\bar{X})$$

equals to

$$SymTE = I((C, X); \overline{(C, X)}) - I(C; X|\overline{(C, X)}) + I(C, X) - I(X_i, \bar{X}) - I(C, \bar{C}) \,,$$

where we used a notation for past of the joint pair $(\bar{C}, \bar{X}_i) = \overline{(C, X_i)}$

Using the relation

$$I(X; \bar{C}|\bar{X}) = H(X|\bar{X}) - H(X|\bar{X}\bar{C}) = H(X|\bar{X}) - H(X) + H(X) - H(X|\bar{X}\bar{C}) = I(X; \bar{C}\bar{X}) - I(X, \bar{X})$$

and similarily

$$I(C; \bar{X}|\bar{C}) = I(C; \bar{X}\bar{C}) - I(C, \bar{C})$$

We consider a sum of both, let's call it symmetrical TE:

$$\begin{aligned} SymTE =& I(C;\bar{X}|\bar{C}) + I(X;\bar{C}|\bar{X}) = I(C;\bar{X}\bar{C}) - I(C;\bar{C}) + I(X;\bar{X}\bar{C}) - I(X;\bar{X}) \\ =& I(C;\bar{X}\bar{C}) + I(X;\bar{X}\bar{C}) - I(C;\bar{C}) - I(X;\bar{X}) \end{aligned}$$

continuing the derivation

$$\begin{aligned} I(C;X) &= H(C) + H(X) - H(C,X) \\ I(C;\bar{X}\bar{C}) &= H(C) - H(C|\bar{X}\bar{C}) \\ I(X;\bar{X}\bar{C}) &= H(X) - H(X|\bar{X}\bar{C}) \\ I(CX;\bar{X}\bar{C}) = H(C,X) - H(C,X|\bar{X}\bar{C}) &= -I(C,X) + H(C) + H(X) - H(C,X|\bar{X}\bar{C}) \\ = -I(C,X) + H(C) + H(X) - H(C,X|\bar{X}\bar{C}) - H(C|\bar{X}\bar{C}) + H(C|\bar{X}\bar{C}) - H(X|\bar{X}\bar{C}) + H(X|\bar{X}\bar{C}) \\ = -I(C,X) + H(C) - H(C|\bar{X}\bar{C}) + H(X) - H(X|\bar{X}\bar{C}) - H(C,X|\bar{X}\bar{C}) + H(C|\bar{X}\bar{C}) + H(X|\bar{X}\bar{C}) \\ = -I(C,X) + I(C,\bar{X}\bar{C}) + I(X,\bar{X}\bar{C}) + I(C,X|\bar{X}\bar{C}) \end{aligned}$$

this gives general equality:

$$I(C,\bar{X}\bar{C}) + I(X,\bar{X}\bar{C}) = I(CX,\bar{X}\bar{C}) - I(C,X|\bar{X}\bar{C}) + I(C,X)$$

plugging back to SymTE:

$$\begin{aligned} SymTE =& I(C,\bar{X}|\bar{C}) + I(X,\bar{C}|\bar{X}) \\ =& I(C,\bar{X}\bar{C}) + I(X,\bar{X}\bar{C}) - I(C,\bar{C}) - I(X,\bar{X}) \\ =& I(CX,\bar{X}\bar{C}) - I(C,X|\bar{X}\bar{C}) + I(C,X) - I(C,\bar{C}) - I(X,\bar{X}) \end{aligned}$$

## References


1. Granger, C.W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
2. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461. [PubMed]
3. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **2009**, *103*, 238701. [CrossRef] [PubMed]
4. Permuter, H.H.; Kim, Y.H.; Weissman, T. On directed information and gambling. In Proceedings of the 2008 IEEE International Symposium on Information Theory, Toronto, ON, Canada, 6–11 July 2008; pp. 1403–1407.
5. Massey, J. Causality, feedback and directed information. In Proceedings of the International Symposium on Information Theory and Its Applications (ISITA-90), Honolulu HI, USA, 27–30 November 1990; pp. 303–305.
6. Griffiths, T.L.; Kemp, C.; Tenenbaum, J.B. Bayesian models of cognition. In *The Cambridge Handbook of Computational Psychology*; Cambridge University Press: Cambridge, UK, 2008.
7. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **2013**, *36*, 181–204. [CrossRef] [PubMed]
8. Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Lanzarote, Spain, 9–11 April 2018; pp. 938–947.
9. Frenzel, S.; Pompe, B. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.* **2007**, *99*, 204101. [PubMed]
10. Suzuki, T.; Sugiyama, M.; Sese, J.; Kanamori, T. Approximating mutual information by maximum likelihood density ratio estimation. In Proceedings of the New Challenges for Feature Selection in Data Mining and Knowledge Discovery, Antwerp, Belgium, 15 September 2008; pp. 5–20.
11. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841.
12. Ma, J. Estimating transfer entropy via copula entropy. *arXiv* **2019**, arXiv:1910.04375.
13. Belghazi, M.I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. Mine: Mutual information neural estimation. *arXiv* **2018**, arXiv:1801.04062.
14. Zhang, J.; Simeone, O.; Cvetkovic, Z.; Abela, E.; Richardson, M. Itene: Intrinsic transfer entropy neural estimator. *arXiv* **2019**, arXiv:1912.07277.
15. Lizier, J.T. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Front. Robot. AI* **2014**, *1*, 11. [CrossRef]
16. Wang, C.i.; Dubnov, S. Guided music synthesis with variable markov oracle. In Proceedings of the Tenth Artificial Intelligence and Interactive Digital Entertainment Conference, Raleigh, NC, USA, 3–7 October 2014.

17. Wang, C.i.; Hsu, J.; Dubnov, S. Music Pattern Discovery with Variable Markov Oracle: A Unified Approach to Symbolic and Audio Representations. In Proceedings of the International Society for Music Information Retrieval Conference, Málaga, Spain, 26–30 October 2015; pp. 176–182.
18. Gokul, V.; Balakrishnan, G.P.; Dubnov, T.; Dubnov, S. Semantic Interaction with Human Motion Using Query-Based Recombinant Video Synthesis. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 379–382.
19. Lefebvre, A.; Lecroq, T. Compror: On-line lossless data compression with a factor oracle. *Inf. Process. Lett.* **2002**, *83*, 1–6.
20. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J.W.; Radford, A.; Sutskever, I. Jukebox: A generative model for music. *arXiv* **2020**, arXiv:2005.00341.
21. Van Den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
22. Poliner, G.E.; Ellis, D.P. A discriminative model for polyphonic piano transcription. *EURASIP J. Adv. Signal Process.* **2006**, *2007*, 48317. [CrossRef]