Maximum entropy models for generation of expressive music

Simon Moulieras[1] **and Francois Pachet**[1,2]

[1]SONY CSL, Paris, France
[2]Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
[*]simon.moulieras@gmail.com

ABSTRACT

In the context of contemporary monophonic music, expression can be seen as the difference between a musical performance and its symbolic representation, *i.e.* a musical score. In this paper, we show how Maximum Entropy (MaxEnt) models can be used to generate musical expression in order to mimic a human performance. As a training corpus, we had a professional pianist play about 150 melodies of jazz, pop, and latin jazz. The results show a good predictive power, validating the choice of our model. Additionally, we set up a listening test whose results reveal that on average, people significantly prefer the melodies generated by the MaxEnt model than the ones without any expression, or with fully random expression. Furthermore, in some cases, MaxEnt melodies are almost as popular as the human performed ones.

Introduction

Human performances of written music differ in many aspects from a straight acoustical rendering a computer would make. Besides the fact that we would not be even able to play music like a machine, a human performer would give its own interpretation to a musical score, by changing the timing, the loudness, the duration, and the timbre of each note, or even by removing or adding notes (in jazz for example). In fact, a human listener can easily recognize a human performance from a straight computer rendering, without necessarily being a musician. Nevertheless, there is no straightforward mathematical formulation of what is or is not musically expressive.

Musical expression is a fundamental component of music. It can translate the intention of the composer via the indications written on the score, in order to help the performer. These indications can be whether associated to a specific note, or a group of notes (dynamical indications, articulations, ...) or a word corresponding to a specific style, or a specific phrasing (*e.g.* fast swing, ballad, groovy, ...). In the latter case, one can legitimately expect to find non-trivial correlations between some well choosen observables, reflecting the specific musical style, and it is the aim of this paper.

For this purpose we will use a MaxEnt model which belongs to the wide class of probabilistic graphical models, built to capture and mimic both unary and pairwise correlations between variables. MaxEnt models have already been used in music[1] in order to study melodic patterns. Unlike in Markov chains in which the sampling complexity grows exponentially with the size of the patterns, MaxEnt models consistent with pairwise correlations keep this complexity quadratic, independently from the distance between the correlated variables. Importantly, we will use a translation-invariant model for simplicity, *i.e* we will consider that the probability distribution of a variable associated to a particular note depends only on the variables associated to the neighbouring notes, independently from its intrinsic position in the melody. This assumption is equivalent to saying that musical expression consists in local texture, rather than long-range correlations.

Like in other aspects of music such as harmony or rhythm, musical expression has a complex underlying structure, that have been tried to be captured by statistical models, or learning algorithm (see[2,3] on romantic music). Working with expression differs from symbolic music computing, in the fact that it involves continuous variables like loudness, or local tempo modulation that cannot be mathematically treated as categories (like the pitch, for example). Unlike romantic music in which a performer usually continuously modulate the tempo, contemporary western music like jazz, or pop music is often played with a rhythmic section that tend to maintain a constant tempo along the melody. The rhythmic expression is thus contained in the difference between: *i*: the onset of the performed note and its onset on the score, and *ii*: the duration of the performed note and its duration on the score. In the following, we will refer to these observables as microtiming deviations.

1 The Model

1.1 Construction

The Principle of Maximum Entropy (PME)[4] states that given a set of observations, the probability distribution that best model the data is the one that maximizes the entropy. The observations are expressed in terms of averages, or expected values, of one or more quantities. This principle has applications in many domains, such as information theory, biology, statistics, and many more, but comes from statistical physics in which it aimed at connecting macroscopic properties of physical systems to a microscopic description at the atomic or molecular level.

Given a set of observables $F_j = \langle f_j(x) \rangle|_{P(x)}$, $1 \leq j \leq N_c$ on a quantity x, in which $\langle \rangle|_{P(x)}$ represents the expectation value over the probability distribution $P(x)$, the PME allows us to determine the distribution $P(x)$ of maximum entropy, fulfilling the constraints.

The variables we need our model to deal with are:

- metrical position in the bar (discrete)

- onset deviation (continuous)

- duration deviation (continuous)

- loudness (continuous).

Hence, the PME should be applied for each of these variables, and to carefully treat separately continuous variables that will be denoted by y, from discrete ones denoted by x. z will refer to an arbitrary variable, belonging to one or the other of the previous types of variable..

Discrete Variables   If x is a discrete variable, taking values in an alphabet $\{x_i\}_{1 \leq i \leq Q}$ then the maximum entropy probability distribution $P(x_i) = Pr(x = x_i)$ reads:

$$P(x_i) = \frac{1}{\mathscr{Z}(\lambda_1, ..., \lambda_{N_c})} \exp\left(\sum_j \lambda_j f_j(x_i)\right) \tag{1}$$

where

$$\mathscr{Z}(\lambda_1, ..., \lambda_{N_c}) = \sum_{i=1}^{Q} \exp\left(\sum_j \lambda_j f_j(x_i)\right). \tag{2}$$

Continuous Variables   On the other hand, if y is a real valued variable, one can show that the probability density function $p(y)$ reads

$$p(y) = \frac{1}{\mathscr{Z}(\lambda_1, ..., \lambda_{N_c})} \exp\left(\sum_j \lambda_j f_j(y)\right) \tag{3}$$

where

$$\mathscr{Z}(\lambda_1, ..., \lambda_{N_c}) = \int_{p(y)>0} \exp\left(\sum_j \lambda_j f_j(y)\right) dy. \tag{4}$$

In both cases, $\mathscr{Z}(\lambda_1, ..., \lambda_{N_c})$ is called partition function and consists in a normalization constant.

We consider now a sequence of N notes, each of which carrying N_{cont} continuous variables(loudness, microtiming deviations) and N_{disc} discrete $X_{v_{disc}}(n)|_{1 \leq n \leq N}$ variables (metrical position in the bar). Transversally, we can see the same sequence as $N_{voices} = N_{cont} + N_{disc}$ sequences of N elements, all of them being whether discrete or continuous numbers. In this view, we will talk about a superposition of voices (the horizontal blue rectangles on Fig. 1), whereas notes are represented vertically (vertical red rectangles).

Let us denote Y_v and X_v respectively a continuous and a discrete variable belonging to the v-th voice, and $Y_v(n)$ and $X_v(n)$ respectively their values at the n-th note. For each voice v, there is a MaxEnt model given by the probability distribution of central variable $Z_v(n)$, conditionned on the values of the neighbouring variables. We define the neighbourhood $\partial Z_v(n)$ of a variable Z_v of the n-th note by itself, the variables surrounding $Z_v(n)$ on the same voice ($Z_v(n \pm k)$ with $1 \leq k \leq K_{\max}^{\text{hor}}$),
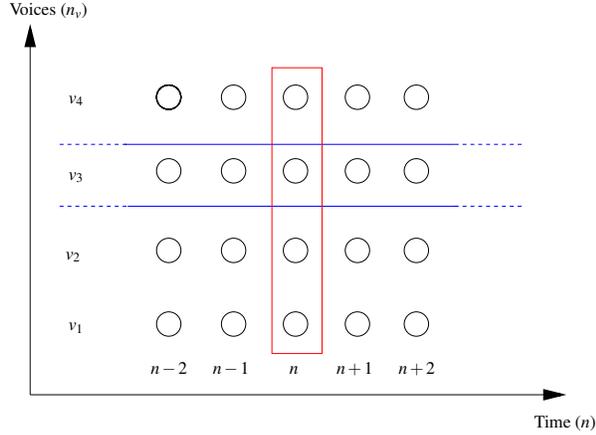
Figure 1. **Schematic view of a sequence.** Graphical representation of a musical sequence of individual notes (red rectangle). Each line represent a voice (blue rectangle), or a specific observable to be involved in the model: microtiming deviations, loudness, metrical position in the bar, etc...
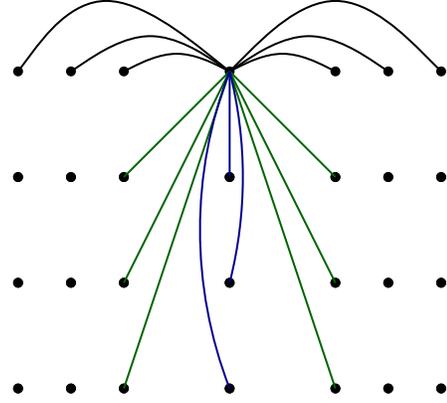

Figure 2. **The Graph Representation.** Graph representing the binary connections (J_v) involved in a model associated to the upper voice v. Connections can be horizontal (black edges, associated with the same voice's variables), vertical (blue edges, associated to different voice's variables but for the same note), or diagonal (green edges, associated to different voice's variables but for different notes).

the variables on the same vertical line ($Z_{v'}(n)$ with $v' \neq v$), and the variables diagonally connected to it ($Z_{v'}(n \pm k)$ with $1 \leq k \leq K_{\max}^{\text{diag}}$). For example, the neighbourhood of a variable is represented on Figure 2, with $K_{\max}^{\text{hor}} = 3$ and $K_{\max}^{\text{diag}} = 1$. The expression of the MaxEnt probability (or probability density, for continuous variables) distribution can be then written as follows:

$$P_v(Z_v|\partial Z_v) = \frac{1}{\mathscr{Z}(\partial Z_v)} \exp\left(-\mathscr{H}_v(Z_v, \partial Z_v)\right) \tag{5}$$

$$\mathscr{H}(Z_v, \partial Z_v) = h_v(Z_v) + \sum_{Z' \in \partial Z_v} J_{Z,Z'} \tag{6}$$

where z correspond to the value of variable Z. In the previous expressions 5 and 6,, the $J_{Z,Z'}$'s are called the couplings (or interaction energy) between two variables Z and Z', and the h_v's the local fields (or bias). Importantly we consider here only symmetric couplings $J_{Z,Z'} = J_{Z',Z}$. This terminology comes from statistical physics in which a hamiltonian $\mathscr{H}(Z_v(n), \partial Z_v(n))$ represents an energy associated to a specific configuration of $Z_v(n)$ and $\partial Z_v(n)$. Depending on the nature of the concerned variables Z and Z', $J_{Z,Z'}$ can whether be a real number, a vector or a matrix.

$$Pr(X_v(n) = x_i|\partial X_v(n)) = P_i(X_v(n)|\partial X_v(n)) = \frac{1}{\mathscr{Z}(\partial X_v(n))} \exp\left(-h_v(x_i) - \sum_{Z \in \partial X_v(n)} J_{X_v,Z}(x_i, z)\right) \tag{7}$$

$$Pr(y < Y_v(n) < y + dy|\partial Y_v(n)) = p_v(y|\partial y_v(n))dy = \frac{1}{\mathscr{Z}(\partial Y_v(n))} \exp\left(-h_v y - \sum_{Z \in \partial Y_v(n)} J_{X_v,Z}(y, z)\right) dy \tag{8}$$

Importantly, due to the choice of observables we made, the self-interaction term $J_{Y,Y}$ for continuous variables Y is proportional to y^2, taking into account the standard deviation 16, while it is zero for integer variables $J_{X,X} = 0$. The rest of the couplings to the concerned continuous variable Y contribute to a linear term (proportional to y). Consequently, the conditional probability density $p_v(y|\partial y_v(n))$ is always a gaussian.

1.2 Learning the parameters

As mentionned in the introduction, we chose to focus on a *translation-invariant* model for simplicity, i.e., for each voice, the elementary module carrying the parameters of the model, represented on figure 2, is translated horizontally for every note of the voice. Every possible position of the graph on the sequence, gives rise to a sample to be used in the inference of the parameters.

Pseudo log-likelihood   We use a *Maximum Likelihood Estimation* (MLE) of the model parameters, i.e., we want to optimize the parameters of a model P_v such that the probability to generate the dataset $\mathscr{D}_v$ associated to the voice v using P_v is maximal. The pseudo log-likelihood approximation,[5,6] consists in substituting the probability of the whole sequence $\mathbf{z} = \{Z_v(n)\}_{1 \leq n \leq N}$ by the product of the conditional probabilities of each sample $\mathbf{s_v} = \{s_v(m)\}_{1 \leq m \leq M}$ where $M = N - 2\max(K_{\max}^{\text{hor}}, K_{\max}^{\text{diag}})$ is the number of samples of the sequence, conditionned on the values of the neighbouring variables. The resulting negative pseudo log-likelihood associated to voice v, $\mathscr{L}_v$ reads:

$$\mathscr{L}_v(\{J_v\}, h_v | \mathscr{D}_v) = -\frac{1}{M} \sum_{m=1}^{M} \log P_v(z_v^m | \partial z_v^m) \tag{9}$$

Since all the couplings are symmetric, the vertical and the diagonal ones are involved in two different models, corresponding to the two different voices. The determination of the N_{voices} models are then entangled and cannot be done sequentially, which lead us to minimize the sum of the negative log-likelihoods over all the voices, consistently with the pseudo log-likelihood approximation.

$$\mathscr{L}(\{\{J_v\}, h_v\}_{1 \leq v \leq N_{\text{voices}}} | \mathscr{D}) = \sum_v \mathscr{L}_v(\{J_v\}, h_v | \mathscr{D}_v) \tag{10}$$

where $\mathscr{D}$ represents the whole dataset $\bigcup_v \mathscr{D}_v$. The minimization of $\mathscr{L}$ in **??** does not generally imply the minimization of the individual $\mathscr{L}_v$, however it gives a good approximation of it while the dataset possesses an overall consistency.[6]

The Corpus   The corpus consists in a series of midi melodies recorded twice[1] by a professional piano player, with a click at a selected tempo and a very neutral accompaniement (chords played on the downbeats). We chose about 172 tunes from the LeadSheet DataBase[7] divided in 5 sub-corpora so to have a balance of expression styles, namely swing, latin, ballad, pop, and groove. The full list of tunes can be found in appendix 3.2. For each tune, the pianist was asked to play the melody freely, without removing any note nor adding any extra note. Of course, being familiar with almost all the tunes, he respected their musical context, and their style conventions: swing phrasing for swing, etc ...

The dataset is formed as follows: the first voice corresponds to the rhythmic score: it is described by the metrical position of a note in its bar, which is encoded in an integer. The normalized midi onset deviation δo, the normalized midi duration deviation δd, and the reduced midi velocity δv (loudness) were then measured for each note, and constitute the continuous voices :

$$\delta o \quad = \quad \frac{\text{onset}_{\text{perf}} - \text{onset}_{\text{score}}}{\text{duration}_{\text{score}}} \tag{11}$$

$$\delta d \quad = \quad \frac{\text{duration}_{\text{perf}} - \text{duration}_{\text{score}}}{\text{duration}_{\text{score}}} \tag{12}$$

$$\delta v \quad = \quad \frac{\text{velocity}_{\text{perf}}}{127} \tag{13}$$

Each sub-corpus was used to train a style model.

1.3 Generating sequences

Once the parameters have been infered, one can generate sequences by sampling from distribution (5) using the Metropolis Algorithm.[8] Every variable of the sequence is initialized randomly. Then, the following step is repeated: randomly select a variable, compute its probability conditioned by its neighbours, given by eq. (7) and eq. (8), and draw a new value from this probability distribution. A stationnary regime is reached after a number of Monte Carlo steps of the order of a few times the number of variables n_{var} (in practice $\approx 10 n_{\text{var}}$).

2 Results

2.1 Quantitative validation

In order to test the capacity of our model to capture the structure of correlations in the learning corpus,one can generate a long sequence (10000 notes), compute both unary and binary correlations, and compare them to the corpus'. On figure 3, one can see the a good agreement for high frequencies (more or less $\geq 10^{-2}$), meaning that patterns sufficiently represented in the

---

[1] one for the training, the other for the validation tests.

corpus are well captured by the model. In our example we took $K_{max}^{hor} = 3$ and $K_{max}^{diag} = 1$, and trained the model on the swing sub-corpus. These values are a good compromise between too many parameters, and a good efficiency in capturing correlations.

Two quantities can be computed along the generation process, in order to monitor the level of convergence of the sequence: The negative log-likelihood, and the distance between the generated sequence correlations, and the corpus' correlations. As we can see on fig. 4, both quantities decrease macroscopically before stabilizing to a stationnary regime, made of typical sequences. We used the same test sequence and the same model (swing) in order to exhibit the convergence properties of the Metropolis procedure.

For each musical style, similar features are observed. A more instructive character of our models is its ability to predict the value of a variable in a sequence that has been played by our performer. This is the role of the test corpus.
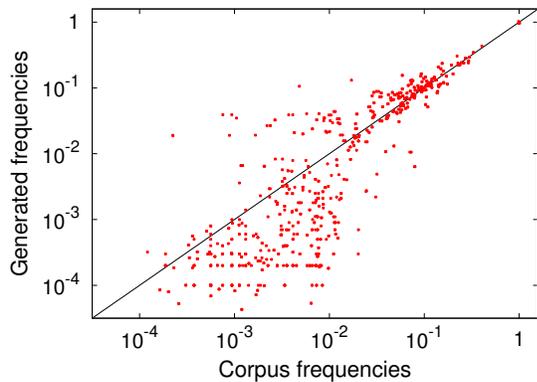

Figure 3. Swing model VS Corpus frequencies: Scatter plot showing a good agreement between the generated sequence and the corpus. $K_{max}^{hor} = 3$ and $K_{max}^{diag} = 1$.
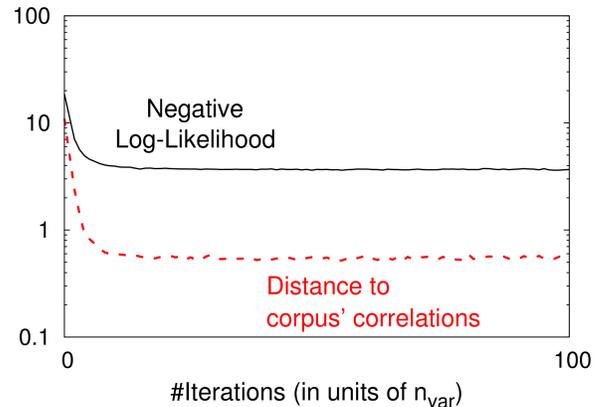

Figure 4. Convergence of the generation: Convergence appears when both quantities remain almost constant. In practice, about $10\,n_{var}$ iterations are needed.

Predictive power: We can perform a leave-one-out cross-validation and obtain good results. Let us remind that the coefficient of determination R^2 measures the predictive power as follows: $R^2 \in [0, 1]$, 0 representing no improved prediction with respect to an empirical gaussian random variable, 1 representing a perfect prediction.

Model Name	Onset deviation	Duration deviation	Loudness
Swing	0.44	0.33	0.24
Pop	0.27	0.09	0.23
Latin	0.33	0.14	0.15
Groove	0.34	0.11	0.21
Ballad	0.32	0.26	0.20

Figure 5. Predictive power of the different models averaged on the test sub-corpora. For each tune of each musical style, we perform a leave-one-out cross-validation procedure, compute the coefficient of determination R^2, and give here its average.

Generally speaking, the onset deviation seems to be better modeled, whereas the loudness gets worse results than the other observables. Note as well that the swing model has, a better predictive power for any observable, than any other model. We will enter into more details in the discussion section 3.

2.2 Perceptive validation

In the previous section, we have shown different mathematical criteria which provide the validation of the structure and the implementation of the MaxEnt models for musical expression. To be consistent with our goal, it is necessary to perform a perceptive validation, and this is how we proceeded:

Protocol: We set up a pairwise comparison test of 4 different versions of a melody: our baseline version called "straight" (with zero microtiming deviations and a constant loudness), a version played by the professional performer (called "human"), a version generated by our models (called "MaxEnt") and a control version, called "random", where microtiming deviations and loudness were sampled with an empirical gaussian distribution[2]. Every participant was asked to listen to two different versions of the melody for each of the 6 tunes we picked. The pair of versions were randomly picked so that after an important number of participations, every version had been compared to every other version the same number of times. The participants were then asked to answer two questions:

- What version did you prefer ?

- What version did you find more expressive ?

For equity purposes, all four versions were rendered with the same method, and with the same average loudness value. The random and MaxEnt versions were generated with a constrained Metropolis algorithm,*i.e.* performing the Metropolis algorithm in which the voice concerning the rhythmic score is fixed to the score's. Furthermore, the accompaniement was rendered with another soundfont, so it could not be confused with the melody. Finally, a participant could do the test only once.

The following table shows the melodies we choose, with the associated MaxEnt model:

Model	Melody	Author
Swing	Donna Lee	C. Parker
	Blues for Alice	
	Bemsha Swing	T. Monk & D. Best
Ballad	Central Park West	J. Coltrane
Latin	Mas que nada	J. Ben Jor
	Meditation	A. C. Jobim

Figure 6. List of the melodies we used for the listening test, with their associated MaxEnt model.

A wide communication was done in order to have as many people as possible do the test. We did not aim at selecting any specific category of people. The website hosting the test has been active for 6 weeks during which it collected 244 participations. The musical samples are available at http://ns2292021.ovh.net:3000/.

Intepretation of the results: In order to give a general ranking over the four versions, we used a so called Bradley-Terry model[9,10] which infers potential β_i for each version i consistent with a probabilistic model giving the probability that version i "beats" version j $(i > j)$:

$$P(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} \tag{14}$$

Denoting $n_{i,j}$ the number of votes $i > j$, we use $P(i > j) = \frac{n_{i,j}}{n_{i,j} + n_{j,i}}$ as input of the inference algorithm, we impose the potential associated to the straight version to be zero. The results are plotted on figure 7.

Note that any set of values of $P(i > j)$ does not necessarily lead to a fully ordered set of β. In the four cases shown on figure 7, the algorithm convergences very quickly, ensuring that the votes are very well fitted by the Bradley-Terry model.

The overall trend puts the human performance in the first position question-wise. The MaxEnt melodies are secondly-prefered, with an significant margin with respect to the other versions. This is particularly the case for the swing, in which there are some strongly defined phrasing codes: delaying the onset and stressing every even eighth of a bar (upbeats), and increasing the duration of every odd eighth of a bar (downbeats). This pattern is very typical, easy to capture by our MaxEnt model, and strongly recognizable by a human ear. On the other hand, expressive phrasing in ballads or a bossa nova tunes is much less stereotyped, and have an important versatiliy, observed in the data. In particular for slow tempo melodies, a slightly too stressed, or slightly too delayed or anticipated note can sound very artificial. It is very likely that the latter reason explains the important differences between the swing results, with the overall trend.

To the question "which melody to you find more expressive?", we observe a different tendency: There is a clear domination of the human performance and a clear inferiority of the straight version. In between, both generated versions get comparable scores. It can be explained by the fact that the extreme versions are very recognizable, and neither random nor MaxEnt models

---

[2]The same distribution corresponding to zero cross validation coefficient of determination in the previous section
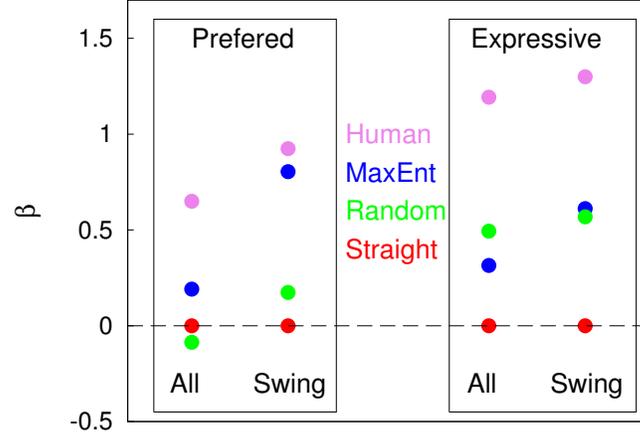
Figure 7. Results of the listening test:. The left (resp. right) panel shows the version potentials for the first (resp. second) question, *i.e.* about the prefered (resp. more expressive) version. In both cases, we observed an important difference between the results coming from the swing tunes, and the other ones, and this is the reason for which we show two columns, one for the 3 swing tunes, and another for the whole test.

success in generating "as much expressivity" as the one perceived in a human performance. In the title of the question, the notion of "more expressive" is confusing, and can be understood as "which version differs the most from the straight melody?". In this sense, the results are consistent with the constructions of the versions, but not very instructive from a perceptive point of view.

3 Conclusion & perspectives

We proposed a Maximum Entropy model which captures pairwise correlations between rhythmic score, micro timing and dynamics observables in musical sequences. The model is used to generate new expression sequences that mimic the interpretation of a musician, given musical style. The structure of this model (see Fig. 2) leads to the replication of complex patterns, despite the pairwise nature of the information used. The pairwise correlations have Moreover, our model deals with both continuous and categorical variables, remaining in the context of Maximum Entropy models.

We performed two different validation processes, quantitative and perceptive, whose conclusions are consistent: MaxEnt models success to capture the local texture of different expression styles. The results are very satisfying in particular for a stereotyped style like swing, whereas they are encouraging for more complex styles, like ballad, or latin jazz.

One of the most challenging perspective is to allow a model of expression to add or remove notes. This model would be very meaningful in from a musical point of view, but would also cost us a deep modification of the modelization. One should also take into account harmony and pitches, keeping a tractable sample complexity thanks to the MaxEnt model.

Appendices

3.1 Appendix: Model details

In this appendix, we expose the target observables that we want to reproduce. First, the statistical mean and standard deviation of a variable Y_v associated to a continuous voice v:

$$f_{\text{mean}}^{\text{cont}}(Y_v) \quad = \quad \frac{1}{N} \sum_{n=1}^{N} Y_v(n) = \bar{Y}_v \tag{15}$$

$$f_{\text{sd}}^{\text{cont}}(Y_v) \quad = \quad \sqrt{\frac{1}{N} \sum_{n=1}^{N} (Y_v(n) - \bar{Y}_v)^2}. \tag{16}$$

Similarly, for discrete variables, we have the frequency of x_i:

$$f_{\text{mean}}^{\text{disc}}(X_v)[x_i] = \frac{1}{N}\sum_{n=1}^{N}\delta_{X_v(n),x_i}, \tag{17}$$

where $\delta_{..}$ is the Kronecker symbol.

Now, we add pairwise correlations between the same variable (on the same voice v) separated by k notes are called horizontal correlations, and read :

$$f_{\text{hor},k}^{\text{cont}}(Y_v) \quad = \quad \frac{1}{N-k}\sum_{n=1}^{N-k}Y_v(n)\,Y_v(n+k) \tag{18}$$

$$f_{\text{hor},k}^{\text{disc}}(X_v)[x_i,x_j] \quad = \quad \frac{1}{N-k}\sum_{n=1}^{N-k}\delta_{X_v(n),x_i}\,\delta_{X_v(n+k),x_j} \tag{19}$$

respectively for a continuous and a discrete variable. Following the same graphical interpretation, vertical correlations exist between variables belonging to voices v_a and v_b, and are defined by :

$$f_{\text{vert}}^{\text{cc}}(Y_{v_a},Y_{v_b}) \quad = \quad \frac{1}{N}\sum_{n=1}^{N}Y_{v_a}(n)\,Y_{v_b}(n) \tag{20}$$

$$f_{\text{vert}}^{\text{dd}}(X_{v_a},X_{v_b})[x_i,x_j] \quad = \quad \frac{1}{N}\sum_{n=1}^{N}\delta_{X_{v_a}(n),x_i}\,\delta_{X_{v_b}(n),x_j} \tag{21}$$

$$f_{\text{vert}}^{\text{cd}}(Y_{v_a},X_{v_b})[x_j] \quad = \quad \frac{1}{N}\sum_{n=1}^{N}Y_{v_a}(n)\,\delta_{X_{v_b}(n),x_j} \tag{22}$$

$$f_{\text{vert}}^{\text{dc}}(X_{v_a},Y_{v_b})[x_i] \quad = \quad \frac{1}{N}\sum_{n=1}^{N}\delta_{X_{v_a}(n),x_i}\,Y_{v_b}(n) \tag{23}$$

respectively for continuous-continuous, discrete-discrete, or continuous-discrete variable pairs. Finally, the diagonal correlations can be written very similarly :

$$f_{\text{diag},k}^{\text{cc}}(Y_{v_a},Y_{v_b}) \quad = \quad \frac{1}{N-k}\sum_{n=1}^{N-k}Y_{v_a}(n)\,Y_{v_b}(n+k) \tag{24}$$

$$f_{\text{diag},k}^{\text{dd}}(X_{v_a},X_{v_b})[x_i,x_j] \quad = \quad \frac{1}{N-k}\sum_{n=1}^{N-k}\delta_{X_{v_a}(n),x_i}\,\delta_{X_{v_b}(n+k),x_j} \tag{25}$$

$$f_{\text{diag},k}^{\text{cd}}(Y_{v_a},X_{v_b})[x_j] \quad = \quad \frac{1}{N-k}\sum_{n=1}^{N-k}Y_{v_a}(n)\,\delta_{X_{v_b}(n+k),x_j} \tag{26}$$

$$f_{\text{diag},k}^{\text{dc}}(X_{v_a},Y_{v_b})[x_i] \quad = \quad \frac{1}{N-k}\sum_{n=1}^{N-k}\delta_{X_{v_a}(n),x_i}\,Y_{v_b}(n+k). \tag{27}$$

3.2 Appendix: Corpus

Yesterday by The Beatles
Billie Jean by Michael Jackson
Eleanor Rigby by The Beatles
Happy by Pharrell Williams
All My Loving by The Beatles
Eight Days A Week by The Beatles
Michelle by The Beatles
And I Love Her by The Beatles
Folsom Prison Blues by Johnny Cash
Fields Of Gold by Sting

So What by Miles Davis
Old Man by Neil Young
Feel by Robbie William
Moondance by Van Morrison
Sir Duke by Stevie Wonder
I Shot The Sheriff by Bob Marley
Flamenco Sketches by Miles Davis
Hello by Lionel Richie
Just The Way You Are by Billy Joel
In The Mood by Glenn Miller
You Are The Sunshine Of My Life by Stevie Wonder
On The Road Again by Willie Nelson
Lively Up Yourself by Bob Marley
Watermelon Man by Herbie Hancock
Both Sides Now by Joni Mitchell
Celebration by Kool & the Gang
Naima by John Coltrane
Giant Steps by John Coltrane
Blue Train by John Coltrane
My Life by Billy Joel
Cantaloupe Island by Herbie Hancock
Chameleon by Herbie Hancock
Goodbye Pork Pie Hat by Charles Mingus
Cousin Mary by John Coltrane
Y.M.C.A by Village People
I Feel Good by James Brown
Jeru by Miles Davis
Maiden Voyage by Herbie Hancock
Every Breath You Take by Sting
Sunny by Bobby Hebb
Strangers Like me by Phil Collins
In A Sentimental Mood by Duke Ellington
Countdown by John Coltrane
Spiral by John Coltrane
Self Portrait In Three Colors by Charles Mingus
Wave by Antonio Carlos Jobim
Locomotion by John Coltrane
Caravan by Duke Ellington & Juan Tizol
Boogie Stop Shuffle by Charles Mingus
Milestones by Miles Davis
Blue Monk by Thelonious Monk
Can't Smile Without You by Barry Manilow
Nuages by Django Reinhardt & Jacques Larue
Epistrophy by Thelonious Monk
Song For My Father by Horace Silver
Fables Of Faubus by Charles Mingus
Lazybird by John Coltrane
Son Of Mr. Green Genes by Frank Zappa
Too High by Stevie Wonder
Mood Indigo by Duke Ellington
Continuum by Jaco Pastorius
Jelly Roll by Charles Mingus
Pussy Cat Dues by Charles Mingus
Part-Time Lover by Stevie Wonder
Now's The Time by Charlie Parker

All in Love Is Fair by Stevie Wonder
Bird Calls by Charles Mingus
Acknowledgement (Part 1 of A LOVE SUPREME) by John Coltrane
Another Star by Stevie Wonder
Strode Rode by Sonny Rollins
Chega de Saudade (No More Blues) by Antonio Carlos Jobim
Fine And Mellow (My Man Don't Leave Me) by null
When You Got A Good Friend by Robert Johnson
Triste by Antonio Carlos Jobim
Resolution (Part 2 of A LOVE SUPREME) by John Coltrane
Evidence by Thelonious Monk
Pannonica by Thelonious Monk
Central Park West by John Coltrane
Sophisticated Lady by Duke Ellington
Bemsha Swing by Thelonious Monk
Brilliant Corners by Thelonious Monk
Solitude by Duke Ellington
Dolphin Dance by Herbie Hancock
That Girl by Stevie Wonder
Four On Six by Wes Montgomery
American Patrol by Glenn Miller
Yardbird Suite by Charlie Parker
In Walked Bud by Thelonious Monk
Long Gone Lonesome Blues by Hank Williams
Equinox by John Coltrane
Stolen Moments by Oliver Nelson
Scrapple From The Apple by Charlie Parker
Little Brown Jug by Glenn Miller
Off Minor by Thelonious Monk
Linus And Lucy by Vince Guaraldi
Infant Eyes by Wayne Shorter
Four by Miles Davis
Impressions by John Coltrane
Spain by Chick Corea
Irene by Caetano Veloso
Love In Vain by Robert Johnson
Confirmation by Charlie Parker
Speak No Evil by Wayne Shorter
Mas Que Nada by Jorge Ben
Thelonious by Thelonious Monk
Witch Hunt by Wayne Shorter
Seven Steps To Heaven by Miles Davis
A Night In Tunisia by Dizzy Gillespie
Chove chuva by Jorge Benjor
Qualquer coisa by Caetano Veloso
Summertime by George Gershwin & Ira Gershwin, Du Bose & Dorothy Heyward
Un Poco Loco by Bud Powell
Friends To Go by Paul Mc Cartney
Sgt Peppers Lonely Hearts Club Band by The Beatles
Butterfly by Herbie Hancock
Donna Lee by Charlie Parker
Queixa by Caetano Veloso
Look To The Sky by Antonio Carlos Jobim
Straight, No Chaser by Thelonious Monk
Ran Kan Kan by Tito Puente

Menino do Rio by Caetano Veloso
Search For Peace by McCoy Tyner
Blues On The Corner by McCoy Tyner
Nefertiti by Miles Davis
Daphne by Django Reinhardt
Crystal Silence by Chick Corea
Petite Fleur (Little Flower) by Sidney Bechet
Meditation by Antonio Carlos Jobim
Blues for Alice by Charlie Parker
Antigua by Antonio Carlos Jobim
Beleza pura by Caetano Veloso
Blue In Green by Bill Evans
Tell Me A Bedtime Story by Herbie Hancock
Speak Like A Child by Herbie Hancock
Luz do Sol by Caetano Veloso
Foi um rio que passou em minha vida by Paulinho da viola
Diminushing by Django Reinhardt
Vivo sonhando by Tom Jobim
Aguas de março by Tom Jobim
Choro by Tom Jobim
Favela by Antonio Carlos Jobim
Falando de amor by Tom Jobim
One Note Samba by Antonio Carlos Jobim
Retrato Em Branco E Preto by Antonio Carlos Jobim
Ligia by Antonio Carlos Jobim
Pais tropical by Jorge Benjor
Aquele abraço by Gilberto Gil
Samba do avião by Tom Jobim
Trilhos Urbanos by Caetano Veloso
Lullaby Of Birdland by George Gershwin
Voce E Linda by Caetano Veloso
Back in Bahia by Gilberto Gil
Samba Cantina by Paul Desmond
I Got You (I Feel Good) by James Brown
Minha Saudade by Joao Donato
Solar by Miles Davis

Acknowledgments

References

1. Sakellariou, J., Tria, F., Loreto, V. & Pachet, F. Maximum entropy model for melodic patterns. In *ICML Workshop on Constructive Machine Learning* (Paris (France), 2015).

2. Chacón, C. E. C. & Grachten, M. An evaluation of score descriptors combined with non-linear models of expressive dynamics in music. In Japkowicz, N. & Matwin, S. (eds.) *Proceedings of the 18th International Conference on Discovery Science (DS 2015)*, Lecture Notes in Artificial Intelligence (Springer, Banff, Canada, 2015).

3. Grachten, M. & Widmer, G. Linear basis models for prediction and analysis of musical expression. *Journal of New Music Research* **41**, 311–322 (2012).

4. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).

5. Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. High-dimensional ising model selection using l1-regularized logistic regression. *Ann. Statist.* **38**, 1287–1319 (2010).

6. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. *Phys. Rev. E* **87**, 012707 (2013). URL http://link.aps.org/doi/10.1103/PhysRevE.87.012707.

7. Pachet, J., F. Suzda & Martín, D. A comprehensive online database of machine-readable lead sheets for jazz standards. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 275–280 (2013). URL http://lsdb.flow-machines.com/.

8. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

9. Ralph Allan Bradley, M. E. T. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39**, 324–345 (1952). URL http://www.jstor.org/stable/2334029.

10. Zermelo, E. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **29**, 436–460 (1929). URL http://dx.doi.org/10.1007/BF01180541.