# *¿El Caballo Viejo?* Latin Genre Recognition with Deep Learning and Spectral Periodicity

Bob L. Sturm[1], Corey Kereliuk[2], and Jan Larsen[2]

[1] School of Electronic Engineering and Computer Science
Queen Mary University of London, Mile End Road London E1 4NS UK
b.sturm@qmul.ac.uk
[2] DTU Compute, Technical University of Denmark, Richard Petersens Plads, B324,
2800 Kgs. Lyngby, Denmark {cmke,janla}@dtu.dk

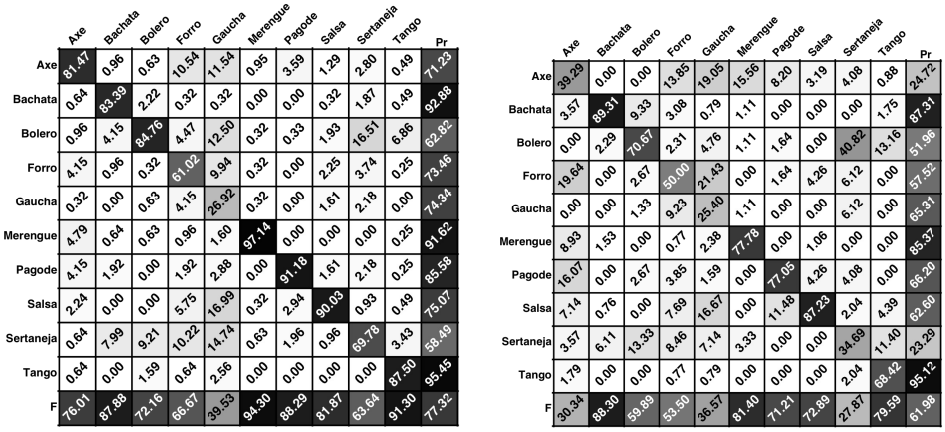
**Abstract.** The "winning" system in the 2013 MIREX Latin Genre Classification Task was a deep neural network trained with simple features. An explanation for its winning performance has yet to be found. In previous work, we built similar systems using the *BALLROOM* music dataset, and found their performances to be greatly affected by slightly changing the tempo of the music of a test recording. In the MIREX task, however, systems are trained and tested using the *Latin Music Dataset (LMD)*, which is 4.5 times larger than *BALLROOM*, and which does not seem to show as strong a relationship between tempo and label as *BALLROOM*. In this paper, we reproduce the "winning" deep learning system using *LMD*, and measure the effects of time dilation on its performance. We find that tempo changes of at most $\pm 6\%$ greatly diminish and improve its performance. Interpreted with the low-level nature of the input features, this supports the conclusion that the system is exploiting some low-level absolute time characteristics to reproduce ground truth in *LMD*.

**Keywords:** machine music listening, genre, deep learning, evaluation

## 1 Introduction

Consider the machine music listening system that "won" the Audio Latin Genre Classification task at MIREX 2013.[3] Among the ten classes in the cleanly labeled Latin Music Database (*LMD*) [14], three systems based on deep learning of spectral periodicity features (DeSPerF) reproduced an average of 77% of the ground truth of each class – more than any of the other systems submitted. Figure 1(a) shows the overall figures of merit (FoM) of these three systems. These FoM, being significantly better than from just guessing, leads one to believe that these systems have successfully learned to identify a good set of musical characteristics associated with each class in *LMD* that are *general* (common to a class) and *discriminative* (distinguishing one class from another). At the heart of this claim, however, sits a false assumption, not to mention an unjustified confidence in the validity of this evaluation.

[3] http://www.music-ir.org/nema_out/mirex2013/results/act/latin_report/summary.html




(a) MIREX (Overall) (b) Our Reproduction

**Fig. 1.** Figures of merit (FoM, $\times 100$) for the (a) three DeSPerF-based systems in MIREX 2013, and (b) our reproduced DeSPerF-LMD system. Column is "true" class, and row is selected class. Off diagonals are confusions. Precision is the right-most column, F-score is the bottom row, recall is the diagonal, and normalised accuracy (mean recall) is at bottom-right corner.

Despite being studied for more than 15 years, music genre recognition (MGR) still lacks an explicit, specific and reasonable definition [1, 19]. The definition most commonly used is that given implicitly by, or by proxy of, some labeled dataset. Critically, the conclusions drawn about systems trained to reproduce labels in a dataset often belie the artificial and unreasonable assumptions made in creating that dataset, not to mention its specious relationship to genre in the real world [6]. Most of these conclusions also implicitly assume that there are only two possible ways to reproduce the ground truth: by chance or with music intelligence [17,19]. When a system reproduces an amount of ground truth much more than that expected from chance, success is declared, and the line of inquiry stops short of proving the outcome to be a result of *music learning*.

In earlier work [20], we sought to explain the winning performance of DeSPerF-based systems in MIREX 2013. Since we did not have access to *LMD* at that time, we used the *BALLROOM* music dataset [4]: a dataset consisting of short music audio excerpts labeled in seven classes. With a 70/30 train and test set partition of *BALLROOM*, we found that the DeSPerF-based system (DeSPerF-BALLROOM) reproduced an average of 88.8% of the ground truth in each class of the test set. We then showed how DeSPerF-BALLROOM can perform perfectly, or no better than random, by time-stretching the test dataset recordings by at most $\pm 6\%$ – effectively changing music tempo without affecting pitch. Furthermore, we showed how minor tempo changes make DeSPerF-BALLROOM label *the same music* in several different ways. For instance, a waltz with a

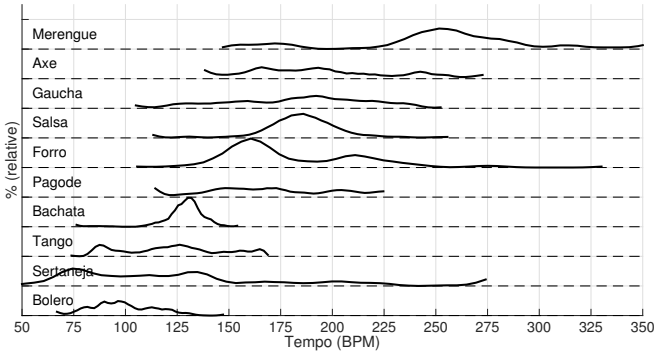



**Fig. 2.** Tempo distributions in *LMD* (smoothed with 3rd-order moving average).

tempo of 87 BPM became a jive at 86 BPM, a rumba at 90 BPM, a samba at 72 BPM, and a cha cha cha at 99 BPM. The explanation for these observations comes from the fact that the labels in *BALLROOM* are highly correlated with the tempo of the excerpts – a characteristic of *BALLROOM* that has been noticed before [4, 7]. Nearest neighbour classification using *only* annotated tempo produces accuracies from 78% [20] to 82.3% [7]. Hence, no matter the meter of the music in the recording, no matter its rhythm, or whether a clave is involved or a bandoneon, a system that can accurately estimate tempo will *appear* quite capable, from being trained and tested in *BALLROOM*, to recognise rhythm, meter, style, instrumentation, and so on.

These results, however, are limited to DeSPerF-BALLROOM and do not explain the "winning" performance of the DeSPerF-based systems in MIREX 2013 (DeSPerF-LMD). Just what has DeSPerF-LMD learned such that it appears able to recognise Merengue with its "crisp, zippy beat, hissed and scratched out on a metal grater quira in jaunty 2/4 time" [14]; Pagode with its "[unpretentious lyrics], focusing on situations from [Brazilian] daily life" [14]; Salsa with its "essential" clave [14]; or Bachata with its standard instrumentation of guitar, maracas and bongos [14]? What has it not learned since it does not appear to recognise Gaucha with its lyrics about "respect for the women, the love for the countryside, the culture and the animals" [14]? A brief look at the characteristics of the labels in *LMD* show that some are musicological and performative, but many are topical, cultural, and geographical, which are of course outside the purview of any artificial algorithm focused exclusively on recorded musical audio. Since DeSPerF-based systems have by design features containing no information about instrumentation or lyrics [12], might DeSPerF-LMD be exploiting a strong correlation between tempo and label in *LMD*?

Recent work [5] suggests that there does not exist a very strong correlation between tempo and label in *LMD*. Figure 2 (created using data collected by Esparza et al. [5]) shows large overlaps in tempo distributions between classes. Table 1 summarises the results from 3-nearest neighbour classification with only estimated tempo, using 10-fold cross validation in *LMD*. Furthermore,

**Table 1.** FoM of 3-NN classification in LMD by tempo (10-fold CV).

| *Class* | Recall | Precision | F-score |
|---|---|---|---|
| *Axe* | 0.1629 | 0.2615 | 0.2008 |
| *Bachata* | 0.6154 | 0.5680 | 0.5908 |
| *Bolero* | 0.4268 | 0.5076 | 0.4637 |
| *Forro* | 0.1538 | 0.2712 | 0.1963 |
| *Gaucha* | 0.3204 | 0.1704 | 0.2225 |
| *Merengue* | 0.7229 | 0.5791 | 0.6431 |
| *Pagode* | 0.1480 | 0.1744 | 0.1601 |
| *Salsa* | 0.2706 | 0.3727 | 0.3136 |
| *Sertaneja* | 0.3156 | 0.4208 | 0.3607 |
| *Tango* | 0.5025 | 0.3764 | 0.4304 |
| Mean | 0.3639 | 0.3702 | 0.3582 |

the size of *LMD* is more than 4.5 times that of *BALLROOM*, and has complete songs instead of 30 second excerpts. Hence, one important question to answer is whether DeSPerF-LMD is as sensitive to "irrelevant" tempo changes as DeSPerF-BALLROOM. If it has a high sensitivity, then DeSPerF-LMD might have learned to exploit absolute temporal characteristics in LMD that are not visible from Fig. 2. If, on the other hand, DeSPerF-LMD is not as sensitive as DeSPerF-BALLROOM, then perhaps DeSPerF-LMD has learned to exploit relative temporal characteristics correlated with the labels of *LMD*, e.g., the rhythmic characteristics to which Pikrakis [12] alludes. Indeed, given the recordings in each of the classes of *LMD* have a far wider variation in tempo than *BALLROOM*, we expect DeSPerF-LMD should be robust to minor changes in tempo as long as the training dataset similarly displays the same variation.

In this work, we seek to answer these questions, which ultimately carry implications for the applications, limitations and improvement of DeSPerF for machine music listening. We begin by reviewing these systems, and then discuss how we create DeSPerF-LMD. We then perform two experiments to determine how time dilation affects the performance of DeSPerF-LMD. We discuss the implications of our results, and propose several avenues for future work.

## 2 DeSPerF-based Systems

### 2.1 The Extraction of SPerF

DeSPerF-based systems combine hand-engineered features – spectral periodicity features (SPerF) – and deep neural networks (DNNs) [12]. A SPerF is generated from an audio extract of 10 seconds. This is broken into 100 ms frames skipped by 5 ms. The first 13 MFCCs [15] are computed for each frame, which produce a *modulation sonogram* $\mathcal{M} = (\mathbf{m}_t : 0 \leq t \leq 10)$, where $\mathbf{m}_t \in \mathbb{R}^{13}$ is a vector of the MFCCs extracted from the frame over time $[t, t + 0.1]$. For *offset* $l \in [1, 4/0.005]$ define the two sequences, $\mathcal{M}_{\text{beg},l} = (\mathbf{m}_t \in \mathcal{M} : 0 \leq t \leq 10 - 0.005l)$ and $\mathcal{M}_{\text{end},l} = (\mathbf{m}_t \in \mathcal{M} : 0.005l \leq t \leq 10)$. $\mathcal{M}_{\text{beg},l}$ are the features starting from the beginning of extract; $\mathcal{M}_{\text{end},l}$ are the features up to its end. The time overlap between features in these two sequences will always be larger than 2 s.

Now, define the *distance* between the sequences for an offset $l$ as

$$d[l] = \frac{\|\mathbf{M}_{\text{beg},l} - \mathbf{M}_{\text{end},l}\|_F}{|\mathcal{M}_{\text{beg},l}|} \tag{1}$$

where the columns of $\mathbf{M}_{\text{beg},l}$ and $\mathbf{M}_{\text{end},l}$ are the sequences $\mathcal{M}_{\text{beg},l}$ and $\mathcal{M}_{\text{end},l}$, and $\|\cdot\|_F$ is the Frobenius norm. The denominator is the number of columns in both matrices. The sequence $d[l]$ is then filtered $y[l] = ((d * h) * h)[l]$, where

$$h[n] = \begin{cases} \frac{1}{n}, & -0.1/0.005 \leq n \leq 0.1/0.005, n \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

and adapting $h[n]$ around the end points of $d[l]$ (shortening its support to a minimum of two). The sequence $y[l]$ then is an approximation of the second derivate of $d[l]$. Finally, a SPerF is created by a non-linear transformation:

$$x[l] = [1 + \exp(-(y[l] - \hat{\mu})/\hat{\sigma})]^{-1} \tag{3}$$

where $\hat{\mu}$ is the mean of $y[l]$ and $\hat{\sigma}$ is its standard deviation. The function of $\hat{\sigma}$ is to remove the influence of energy in the modulation sonograms computed from many audio extracts, thereby making them comparable.

From this derivation, we clearly see that SPerF describe temporal periodicities of modulation sonograms. If the sequence $\mathcal{M}$ is periodic with period $T$ seconds, then the sequence $d[l]$ should be small, and $y[l]$ should be large positive, for all $l \approx kT/0.005$ with $k$ a positive integer. At some of these offsets $x[l]$ will be close to 1. The hope is that $x[l]$ provides insight into musical characteristics such as tempo, meter and rhythm, when they exist over durations of at most 10 seconds. An estimate of a multiple of the tempo can come from a spectral analysis of $x[l]$, i.e., the amplitudes and frequencies of its harmonics. Predicting meter requires approaches that are more heuristic, e.g., deciding on the beat level and then grouping peaks of $x[l]$. Describing rhythm from $x[l]$ should involve even more heuristics, not to mention information SPerF does not contain, e.g., instrumentation. By using SPerF as input to deep learning systems, one hopes that it automatically develops such heuristics meaningful for music listening.

### 2.2 The Construction and Operation of DeSPerF-based Systems

The deep neural network (DNN) used in DeSPerF-based systems specifically use feedforward architectures, whereby the input data is propagated through one or more hidden layers. This forward propagation is achieved via a series of cascaded operations consisting of a matrix multiplication followed by a non-linear function (e.g., a logistic sigmoid or hyperbolic tangent). Since each DNN layer computes a feature representation of the data in the previous layer, the hidden layers are said to compute "features-of-features." The hierarchical nature of DNNs is a commonly cited motivation for choosing to work with them [3,10]. For instance, it might be argued that music rhythm perception is hierarchical in nature, e.g.,

beat-level, measure-level, and so on, which motivates the application of DNNs to recognising rhythmic qualities in recorded music.

Several efficient DNN training techniques have been developed [2, 8, 16]. A DeSPerF-based system employs a DNN trained using a common two-phase process: unsupervised pre-training followed by supervised fine-tuning. Pre-training initializes the network with a 'good' set of weights, which can be critical for achieving learning times that are independent of depth [13]. The pre-training phase is accomplished by greedily training a stack of restricted Boltzmann machines using 1-step contrastive divergence [8]. In the subsequent fine-tuning step, backpropagation is used to adjust the network parameters in order to minimize the expected misclassification error on the labeled training data. The DeSPerF-based systems in MIREX 2013 have five hidden layers with 400 units each.

The final layer of a DeSPerF-based system involves a softmax unit, the output of which can be interpreted as the posterior probability of the classes for an input SPerF. The class of the largest posterior is thus applied to the unlabelled observation. In *LMD*, however, observations are longer than 10 s, and so a single music recording can produce many SPerF. Since the classification problem implicit in *LMD* is to classify whole recordings and not 10 s excerpts, the DeSPerF-based systems in MIREX employs majority vote. In other words, for each SPerF extracted from the same music recording, a vote is recorded for the class of the maximum posterior probability. Once all SPerF have been processed, the class with the most votes is selected.

### 2.3 Evaluating DeSPerF-LMD

Though we have access to *LMD* we do not have access to the specific folds used in this MIREX task. We thus reverse engineer the folds given the results of MIREX 2013,[4] and using the claim that the task employs artist filtering. Table 2.3 shows the number of tracks of each class appearing in each fold. We create an approximately 70/30 train/test partition by combining the numbers in the coloured cells. Our copy of LMD contains 3229 excerpts (1 extra each in Pagode and Sertaneja). We compose the train and test folds using the blocks of artists in each class. Since more than 81% (334/408) of the tracks in LMD Tango are by one artist (Carlos Gardel), we have to violate artist filtering by including 41 excerpts of his in the test set. We use his first 41 excerpts listed by filename.

Figure 1(b) shows the FoM of our DeSPerF-LMD system. Comparison with Fig. 1(a) shows some possible discrepancies. First, the normalised accuracies differ by more than 15 points; however, the FoM in Fig. 1(a) is the overall FoM for three systems tested on the three folds summarised by Table 2.3. In fact, the three normalised accuracies measured in the MIREX 2013 folds for each DeSPerF-LMD system are reported 73.34, 65.81 and 51.75.[5] Hence, our observation of 61.98 is not alarming. With the exception of Bachata, the FoM of our system is worse than those seen in MIREX 2013. We see large differences in

[4] `http://www.music-ir.org/nema_out/mirex2013/results/act/latin_report/files.html`

[5] The fold composition in the MIREX task is problematic. Table 2.3 shows folds 1 and 2 are missing examples of 2 classes, and fold 1 has only one example in another.

**Table 2.** An overview of the composition of the three folds used in the 2013 MIREX Audio Latin Genre Classification Train-test Task. We construct an approximately 70/30 split in each class by combining the shaded numbers of tracks to the test partition.

| \ Fold<br>*Class* \ | 1 | 2 | 3 | Total | Proportion in our test |
|---|---|---|---|---|---|
| *Axe* | 257 | 14 | 42 | 313 | 17.9 % |
| *Bachata* | 1 | 131 | 181 | 313 | 41.9 % |
| *Bolero* | 68 | 172 | 75 | 315 | 23.8 % |
| *Forro* | 183 | 0 | 130 | 313 | 41.5 % |
| *Gaucha* | 0 | 126 | 186 | 312 | 40.4 % |
| *Merengue* | 224 | 80 | 11 | 315 | 28.9 % |
| *Pagode* | 60 | 246 | 0 | 306 | 19.6 % |
| *Salsa* | 75 | 217 | 19 | 311 | 30.2 % |
| *Sertaneja* | 0 | 272 | 49 | 321 | 15.3 % |
| *Tango* | 114 | 0 | 294 | 408 | 27.9 % |
| Totals | 982 | 1258 | 987 | 3227 | 28.7 % |

the recall and precision for Axe, Merengue, Sertaneja, Tango and Bolero. Again, looking over the FoM for the individual systems in MIREX 2013, these are not alarming. Of all DeSPerF-LMD systems tested in MIREX 2013, the one built using folds 1 and 2 performed the worst in these classes. For Axe, its recall and precision was 0.43 and 0.23, respectively. For Merengue, these were 0.72 and 0.35; for Sertaneja: 0.43 and 0.15; for Bolero: 0.71 and 0.37; and for Tango: 0.82 and 0.95. Hence, we conclude that our DeSPerF-LMD system is working comparably to those built in MIREX 2013 with respect to their FoM.

## 3 Measuring the Sensitivity to Tempo of DeSPerF-LMD

Given the above results, we now attempt to inflate and deflate its FoM by the method of irrelevant transformation [18] through pitch-preserving time stretching using the RubberBand library.[6] We then attempt to make DeSPerF-LMD apply different labels to the same music by the same transformation.

### 3.1 Inflation and Deflation of FoM

By the same deflation and inflation procedures we applied in [20], we find that DeSPerF-LMD obtains the FoM shown in Fig. 3 with changes of at most ±6% (i.e., a dilation factor 0.94 or 1.06). Comparison with Fig. 1(b) shows severe harm or great benefit to the FoM of DeSPerF-LMD. If we change the tempo by at most ±10%, we find the normalised classification accuracy reaches 0.11 with deflation, and 0.94 with inflation. Figure 4 shows how even for small tempo changes the F-scores for all classes are dramatically affected.

### 3.2 Picking Any Class in LMD

We now randomly select one excerpt of each label from the test set, and attempt to make DeSPerF-LMD classify them in every way. Table 3 shows the

[6] `http://breakfastquay.com/rubberband/`

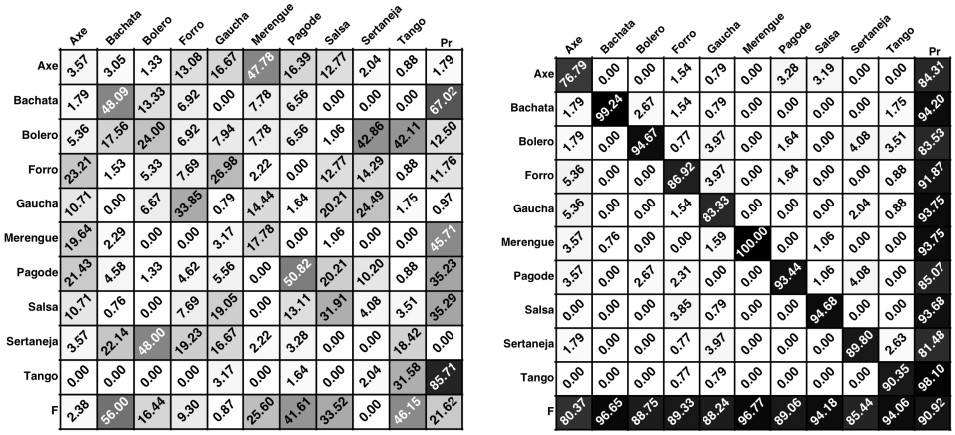



**Fig. 3.** FoM ( $\times 100$) resulting from deflation (left) and inflation (right) for DeSPerF-LMD. The maximum change in tempo here is $\pm 6\%$. Interpretation as in Fig. 1.

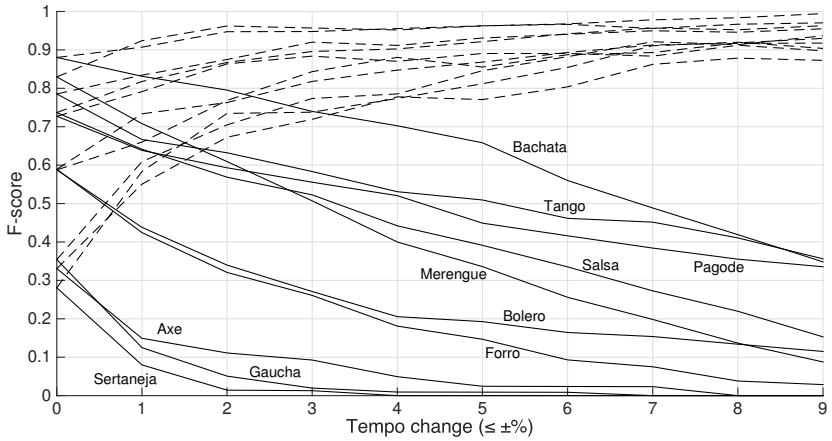



**Fig. 4.** Inflation (dashed) and deflation (solid) of F-score of DeSPerF-LMD in each label of *LMD* as a function of the maximum change in tempo. We label only the deflation, but inflation for each class begins from the same F-score.

new tempo of each track (using data from [5]), and the resulting classifications.[7] From Fig. 1(b) we see that Bachata receives the highest recall and F-score, and second highest precision. DeSPerF-LMD classifies the Bachata tracks in 6 categories, with the largest change of 0.79. It classifies as Bachata six not-Bachata tracks, with the largest change of 1.23. In Merengue, DeSPerF-LMD has the second highest F-score, and third highest precision. It labels the Merengue excerpt

[7] Audition this table at `http://www.eecs.qmul.ac.uk/~sturm/research/DeSPerFtable2/exp.html`

**Table 3.** *LMD* tracks (left column, and red circles in Fig. **??**) are classified in a number of different ways by time-stretching. Resulting estimated tempi are shown.

| **Label** *Title* (tempo) \ *Class* | *Axe* | *Bachata* | *Bolero* | *Forro* | *Gaucha* | *Merengue* | *Pagode* | *Salsa* | *Sertaneja* | *Tango* |
|---|---|---|---|---|---|---|---|---|---|---|
| **Axe** *Batom Na Cueca, Arrasta* (168.36) | 168.36 | 141.48 | 134.69 | 160.34 | 177.22 | 155.89 | 181.03 | 154.46 | 157.34 | |
| **Bachata** *Aventura, Ensename A Olvidar* (124.71) | | 125.97 | 118.77 | 145.01 | 113.37 | | | 164.09 | 124.71 | |
| **Bolero** *Emilio Santiago - Dilema* (101.92) | | 118.52 | 101.92 | 80.89 | | | 141.56 | 113.25 | 102.95 | |
| **Forro** *Trio Forrozao - Ze do Rock* (163.85) | 176.18 | | | 163.85 | 180.05 | | | 195.06 | 142.48 | |
| **Gaucha** *Alma Serrana, O Garrafao* (178.82) | 178.82 | 155.50 | | 173.61 | 177.05 | | 158.25 | 184.35 | | |
| **Merengue** *Ronny Moreno - Matame* (279.49) | 288.14 | 268.74 | | 321.26 | 234.87 | 279.49 | 297.33 | 221.82 | 220.07 | |
| **Pagode** *Grupo Sensacao, Sorrriso de marfin* (142.11) | 175.45 | 129.19 | 122.51 | | | | 142.11 | 194.67 | 124.66 | |
| **Salsa** *Eddie Santiago, Hagamoslo* (168.67) | 168.67 | 165.36 | | 170.37 | 191.67 | | 167.00 | 172.11 | 137.13 | |
| **Sertaneja** *Leandro & Leonardo, Eu Juro* (87.04) | | | 87.04 | | 106.15 | 70.77 | | | 83.70 | 69.64 |
| **Tango** *A Passion For Tango, Milonga de Mis Amores* (112.29) | 155.96 | | 113.43 | 111.18 | 129.07 | | 142.14 | | 112.29 | 118.20 |

eight different ways. The hardest classification to force was Tango, where only one not-Tango track was classified Tango with a change 1.25.

### 3.3 Pick Any Class outside of LMD

We now attempt to make DeSPerF-LMD classify in every way time-stretched versions of the ten music recording excerpts used in [18]. Table 4 shows that we were able to do this for most labels and with minor time-stretching factors.

## 4 Discussion

The results of our experiments show the performance of DeSPerF-LMD to be strongly dependent upon some characteristic related to *absolute time*. Figure 3 shows the normalised accuracy of DeSPerF-LMD drops 40 points or increases 30 with tempo changes of at most $\pm 6\%$. Figure 4 shows that small tempo changes greatly impact the reproduction of ground truth of all labels. Table 3 shows DeSPerF-LMD can be made to classify several *LMD* excerpts in most ways it has learned; and Table 4 shows the same result for music excerpts that are not a part of *LMD*. Though Fig. 1 is evidence that DeSPerF-LMD has certainly learned something about *LMD*, the results of our experiments show that what it has learned may not be of much use when it comes to identifying or discriminating

| **"Original" (tempo)** | *Axe* | *Bachata* | *Bolero* | *Forro* | *Gaucha* | *Merengue* | *Pagode* | *Salsa* | *Sertaneja* | *Tango* |
|---|---|---|---|---|---|---|---|---|---|---|
| Little Richard *Last Year's Race Horse* (82.00) | 96.47 | 110.81 | 82.00 | 79.61 | 78.10 | | 128.12 | 113.89 | 81.19 | 70.09 |
| Rossini *William Tell Overture* (164.00) | 165.66 | 146.43 | 164.00 | 157.69 | 160.78 | 133.33 | 298.18 | 182.22 | 150.46 | 140.17 |
| Willie Nelson *A Horse Called Music* (63.00) | 70.79 | | 68.48 | 66.32 | 56.25 | | 75.00 | 92.65 | 70.00 | 63.00 |
| Simian Mobile Disco *10000 Horses Can't Be Wrong* (130.00) | 128.71 | | | 106.56 | 113.04 | 130.00 | 149.43 | 111.11 | 114.04 | |
| Rubber Bandits *Horse Outside* (114.00) | 110.68 | 121.28 | 109.62 | 142.50 | 112.87 | 114.00 | | 193.22 | 106.54 | |
| Leonard Gaskin *Riders in the Sky* (95.00) | 84.07 | 120.25 | 95.00 | 82.61 | 66.43 | 68.84 | 148.44 | 102.15 | 95.96 | 74.22 |
| Jethro Tull *Heavy Horses* (113.00) | 97.41 | 124.18 | 114.14 | | 113.00 | 221.57 | 137.80 | 166.18 | 108.65 | 125.56 |
| Echo and The Bunnymen *Bring on the Dancing Horses* (120.00) | 118.81 | 127.66 | 104.35 | 146.34 | 114.29 | 120.00 | | 110.09 | 115.38 | |
| Count Prince Miller *Mule Train* (91.00) | 95.79 | 121.33 | 91.00 | 86.67 | 105.81 | 88.35 | | 110.98 | 94.79 | |
| Rolling Stones *Wild Horses* (70.00) | 51.09 | | 71.43 | 54.26 | 75.27 | | | | 70.00 | 68.63 |

**Table 4.** Not-*LMD* tracks (left column) are classified in a number of different ways by time-stretching. Resulting tempi (found manually) are shown.

Latin music genre or style. An impressive precision in Bachata inspires hope that DeSPerF-LMD has automatically learned why something does or does not "sound like" Bachata. By the results in Table 3, DeSPerF-LMD says slightly speeding up the Bolero excerpt makes it sound more like Bachata than Bolero; and slightly slowing down the Bachata excerpt make it sound more like Bolero then Bachata. Table 4 shows how for DeSPerF-LMD the "original" excerpt of the "William Tell Overture" sounds most like Bolero, but slowing it down 11% makes it sound more like Bachata, slowing it down by 15% makes it become Tango, and slowing it down 19% creates Merengue. This is not good behaviour.

In their brief musicological descriptions of the music labels in *LMD*, Silla et al. [14] allude to tempo only twice: Merengue has a "zippy" beat, and Axe is "energetic." Supported by Fig. 2, minor changes in tempo should be insignificant to *LMD*. For the mode of the narrowest distribution (Bachata, 130 BPM), a tempo change of $\pm 6\%$ is a difference of about 8 BPM. For the mode of the Merengue tempo distribution (255 BPM), such a change is a difference of about 15 BPM. Since these intervals are well within the spreads of each distribution, one hopes DeSPerF-LMD would not be so sensitive to these changes. While the input SPerF are by construction intimately connected to absolute time characteristics (Sect. 2.1), the results of our experiments suggest that the deeper features produced by deep learning are sensitive to changes of a characteristic that has minor importance for the designation of a recording of music as any *LMD* label.

From the size of *LMD*, the distribution of tempi of its excerpts, and the fact that the FoM in Fig. 1 are produced using artist filtering, it is hard to believe there to be a specific absolute time characteristic confounded with the labels. In our previous experiments [20], we found the mechanism introducing such a confound into the *BALLROOM* dataset. So, we must discover the cue used by DeSPerF-LMD to produce an illusion of music understanding. An opportunity for this is given in Fig. 3(b) and Fig. 4. Analysing the SPerF extracted from the set of time-stretched test signals inflating these FoM might reveal the cues learned by DeSPerF-LMD. While these are negative results, they are also opportunities to improve assumptions and models, as well as machine music listening systems and approaches to their evaluation. Our work motivates in particular the transformation of SPerF (3) to be time-relative rather than time-absolute, and then to measure the impact of this change by performing the experiments above with the new system. Our work also suggests new ways to evaluate systems submitted to the MIREX LMD task, and in fact any of its train-test tasks.

## 5 Conclusion

DeSPerF-LMD appears to be quite adept at a complex human feat, in spite of the fact that it does not have access to many of the most significant characteristics identifying and distinguishing the labels in *LMD* (e.g., topical, geographical, instrumental). When one claims the only two explanations for such an outcome are either by chance or by music learning, it is easy to see why one would accept the conclusion that the system has learned something general and useful about music. Along with our results in [20], there is however little to support the claim that DeSPerF-based systems trained in *BALLROOM* and in *LMD* have learned anything general about music, meter or rhythm. The story of Clever Hans [11,18] provides a third and much more reasonable explanation: the old horse (*el caballo viejo*) has learned to exploit cues hidden by the lack of control over the independent variables of the evaluation. Once these cues are removed, e.g., giving Clever Hans a private office in which to solve the firm's accounting, or slightly adjusting the tempo of a music recording, the horse reveals its shenanigans.

Speaking more broadly, the 2013 MIREX victory of DeSPerF-LMD, and indeed any victory in the current MIREX train-test tasks, is hollow. When an experimental design lacks an accounting for all independent variables, then one cannot conclude a system has learned to solve some problem implicitly defined by a labeled dataset *no matter how good is its FoM* [17, 19]. A machine music listening system can appear to be solving a complex listening task merely by exploiting irrelevant but confounded factors [18, 20]. "Solutions" will freely masquerade as advancements until evaluation methods are required to possess the relevance and validity to make the desired conclusions. The development of these valid methods is impossible as long as the problem remains undefined; but we have shown in this paper that it is possible to test claims such as: "System X is performing significantly better than random because it has learned something general about music." It just requires thinking outside the stable.

**Acknowledgments** We greatly appreciate Aggelos Pikrakis for making his code available for analysis and testing. CK and JL were supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

## References

1. Aucouturier, J.J., Pachet, F.: Representing music genre: A state of the art. J. New Music Research 32(1), 83–93 (2003)
2. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. Advances in neural information processing systems 19, 153 (2007)
3. Deng, L., Yu, D.: Deep Learning: Methods and Applications. Now Publishers (2014)
4. Dixon, S., Gouyon, F., Widmer, G.: Towards characterisation of music via rhythmic patterns. In: Proc. ISMIR. pp. 509–517 (2004)
5. Esparza, T., Bello, J., Humphrey, E.: From genre classification to rhythm similarity: Computational and musicological insights. J. New Music Research (2014)
6. Frow, J.: Genre. Routledge, New York, NY, USA (2005)
7. Gouyon, F., Dixon, S., Pampalk, E., Widmer, G.: Evaluating rhythmic descriptors for musical genre classification. In: Proc. Audio Eng. Soc. Conf. pp. 196–204 (2004)
8. Hinton, G., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation 18(7), 1527–1554 (2006)
9. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 6(2), 107–116 (Apr 1998)
10. Humphrey, E., Bello, J., LeCun, Y.: Feature learning and deep architectures: New directions for music informatics. J. Intell. Info. Systems 41(3), 461–481 (2013)
11. Pfungst, O.: Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology. Henry Holt, New York (1911)
12. Pikrakis, A.: A deep learning approach to rhythm modeling with applications. In: Proc. Int. Workshop Machine Learning and Music (2013)
13. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. CoRR abs/1312.6120 (2013)
14. Silla, C.N., Koerich, A.L., Kaestner, C.A.A.: The Latin music database. In: Proc. ISMIR (2008)
15. Slaney, M.: Auditory toolbox. Tech. rep., Interval Research Corporation (1998)
16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014)
17. Sturm, B.L.: Classification accuracy is not enough: On the evaluation of music genre recognition systems. J. Intell. Info. Systems 41(3), 371–406 (2013)
18. Sturm, B.L.: A simple method to determine if a music information retrieval system is a "horse". IEEE Trans. Multimedia 16(6), 1636–1644 (2014)
19. Sturm, B.L.: The state of the art ten years after a state of the art: Future research in music information retrieval. J. New Music Research 43(2), 147–172 (2014)
20. Sturm, B.L., Kereliuk, C., Pikrakis, A.: A closer look at deep learning neural networks with low-level spectral periodicity features. In: Proc. Int. Workshop on Cognitive Info. Process. (2014)