Submitted to the Annals of Applied Statistics

SUPPLEMENT TO MARKOV-SWITCHING STATE SPACE MODELS FOR UNCOVERING MUSICAL INTERPRETATION

BY DANIEL J. MCDONALD[1], MICHAEL MCBRIDE[2,*], YUPENG GU[3,†], AND CHRISTOPHER RAPHAEL[3,‡]

[1]*Department of Statistics, University of British Columbia, daniel@stat.ubc.ca*

[2]*Department of Statistics, Indiana University, *michmcbr@iu.edu*

[3]*School of Informatics, Computing, and Engineering, Indiana University, †yupeng.gu@gmail.com; ‡craphael@indiana.edu*

1. Algorithms. For completeness, we include here concise descriptions of the Kalman filter and smoother we employ as inputs to our main algorithm. The filter is given in Algorithm A1.

Algorithm A1 Kalman filter: estimate x_i conditional on $\{y_j\}_{j=1}^{i}$, for all $i = 1, \ldots, n$ and calculate the log likelihood for θ

Input: Y, x_0, P_0, d, T, c, Z, and G
$\ell(\theta) \leftarrow 0$ ▷ Initialize the log-likelihood
for $i = 1$ to n **do**
  $\chi_i \leftarrow d + T x_{i-1|i-1}, \quad P_i \leftarrow Q + T P_{i-1|i-1} T^\top$ ▷ Predict current state
  $\widetilde{y}_i \leftarrow c + Z\chi_i, \quad F_i \leftarrow G + Z P_i Z^\top$ ▷ Predict current observation
  $v_i \leftarrow y_i - \widetilde{y}_i \quad K_i \leftarrow P_i Z^\top F^{-1}$ ▷ Forecast error and Kalman gain
  $x_{i|i} \leftarrow \chi_i + K_i v_i, \quad P_{i|i} \leftarrow P_i - P_i Z^\top K_i$ ▷ Update
  $\ell(\theta) = \ell(\theta) - v_i^\top F^{-1} v_i - \log(|F_i|)$
end for
return $\widetilde{Y} = \{\widetilde{y}_i\}_{i=1}^n$, $\chi = \{\chi_i\}_{i=1}^n$, $\widetilde{X} = \{x_{i|i}\}_{i=1}^n$, $P = \{P_i\}_{i=1}^n$, $\widetilde{P} = \{P_{i|i}\}_{i=1}^n$, $\ell(\theta)$

To incorporate all future observations into these estimates, the Kalman smoother is required. There are many different smoother algorithms tailored for different applications. Algorithm A2, due to Rauch et al. (1965), is often referred to as the classical fixed-interval smoother (Anderson and Moore, 1979). It produces only the unconditional expectations of the hidden state $\widehat{x}_i = \mathbb{E}[x_i \mid y_1, \ldots, y_n]$ for the sake of computational speed. This version is more appropriate for inference in the type of switching models we discuss in the manuscript.

Algorithm A2 Kalman smoother (Rauch-Tung-Striebel): estimate $\widehat{X}$ conditional on Y

Input: χ, $\widetilde{X}$, P, $\widetilde{P}$, T, c, Z.
$i = n$,
$\widehat{x}_n \leftarrow \widetilde{x}_n$,
while $t > 1$ **do**
  $\widehat{y}_i \leftarrow c + Z\widehat{x}_i$, ▷ Predict observation vector
  $e \leftarrow \widehat{x}_i - \chi_i, \quad V \leftarrow P_i^{-1}$,
  $i \leftarrow i - 1$, ▷ Increment
  $\widehat{x}_i = \widetilde{x}_i + \widetilde{P}_i T V e$
end while
return $\widehat{Y} = \{\widehat{y}_i\}_{i=1}^n, \widehat{X} = \{\widehat{x}_i\}_{i=1}^n$

TABLE 1
The factor loadings for principal component analysis of the parameter estimates.

	σ^2_ϵ	μ_{tempo}	μ_{acc}	μ_{stress}	σ^2_{tempo}	p_{11}	p_{12}	p_{31}	p_{13}	p_{21}	p_{32}	p_{22}
PC1	0.07	0.34	-0.66	-0.22	-0.09	0.27	-0.34	0.32	0.03	-0.05	-0.3	-0.05
PC2	0.07	0.00	-0.12	-0.48	-0.02	-0.46	0.13	-0.01	0.17	0.68	0.0	-0.17
PC3	0.64	-0.59	0.16	0.02	-0.03	0.08	-0.13	0.32	0.03	0.05	-0.3	-0.02

2. Principal components. In Section 3.3 of the main document, we plotted the first two principal components along with some notion of groups to gauge the similarities between performances. Table 1 gives the loadings for the first three principal components. We see that the first component picks up information about the first two states, both through μ_{tempo} and μ_{acc} as well as loading onto the probabilities p_{11}, p_{12}, and p_{31}. The second component loads especially onto μ_{stress} but also p_{11} and p_{21}. Finally, the third component loads mainly onto the observation error with smaller contributions from p_{31}.

3. Confidence intervals. Section 3 of the manuscript includes parameter estimates for some of the recordings in our data set. In order to quantify uncertainty and compare the estimates, this section graphically displays all parameter estimates in Figure SM-1. The recordings are sorted in the same order as in Figure 9 in the main document, so some of the conclusions about groupings are readily apparent. The bars indicating measures of uncertainty are derived form the observed Fisher information from the optimization routine. However, it's not entirely clear what these mean. For one, they ignore any uncertainty in the state sequence (see Figure SM-10 some notion of the scale of this uncertainty). They also depend on identifiability, the priors, and the approximation to the posterior. Producing the MAP depends on the approximation accuracy at the MAP, but producing the Hessian needs that as well as the accuracy of about $5p^2$ additional function evaluations. And because we need the inverse, any inaccuracies could explode. The length of the confidence interval for parameter j is given by $4\sqrt{(\widehat{I})^{-1}_{jj}}$ and so these would have roughly 95% coverage. For any parameters that are unidentified, the width of the band is the maximum over all performances for the same parameter. However, short of performing a fully-Bayesian analysis, we would hesitate to attach much certainty to these metrics of uncertainty.

4. Distance matrix from raw data. In Section 3.3 of the manuscript, we present results for grouping performances using the low-dimensional vector of performance specific parameters learned for our model. An alternative approach is to simply use the raw data, in this case, 231 individual note-by-note instantaneous speeds measured in beats per minute. In Figure SM-2 we show the result of this analysis. A comparison between this clustering and that given by our model is discussed in some detail in the manuscript.

5. Plotting performances. Section 3.3 of the manuscript discusses 7 groups of recordings. Figures SM-3 to SM-9 display the note-by-note tempos along with the inferred interpretive decisions for all performances based on this grouping.

The first group (Figure SM-3 indicated as ∘ in Figure 10) corresponds to reasonably staid performances. This group is the largest and corresponds to the block from Cohen to Brailowsky in Figure 9. In this group, the emphasis state is rarely visited with the performer tending to stay in the constant tempo state with periods of slowing down at the ends of phrases. Acceleration is almost never used. Furthermore, these performances have relatively slow average tempos, and not much difference between the A and B sections. Joyce Hatto's recording in Figure 7 is typical of this group

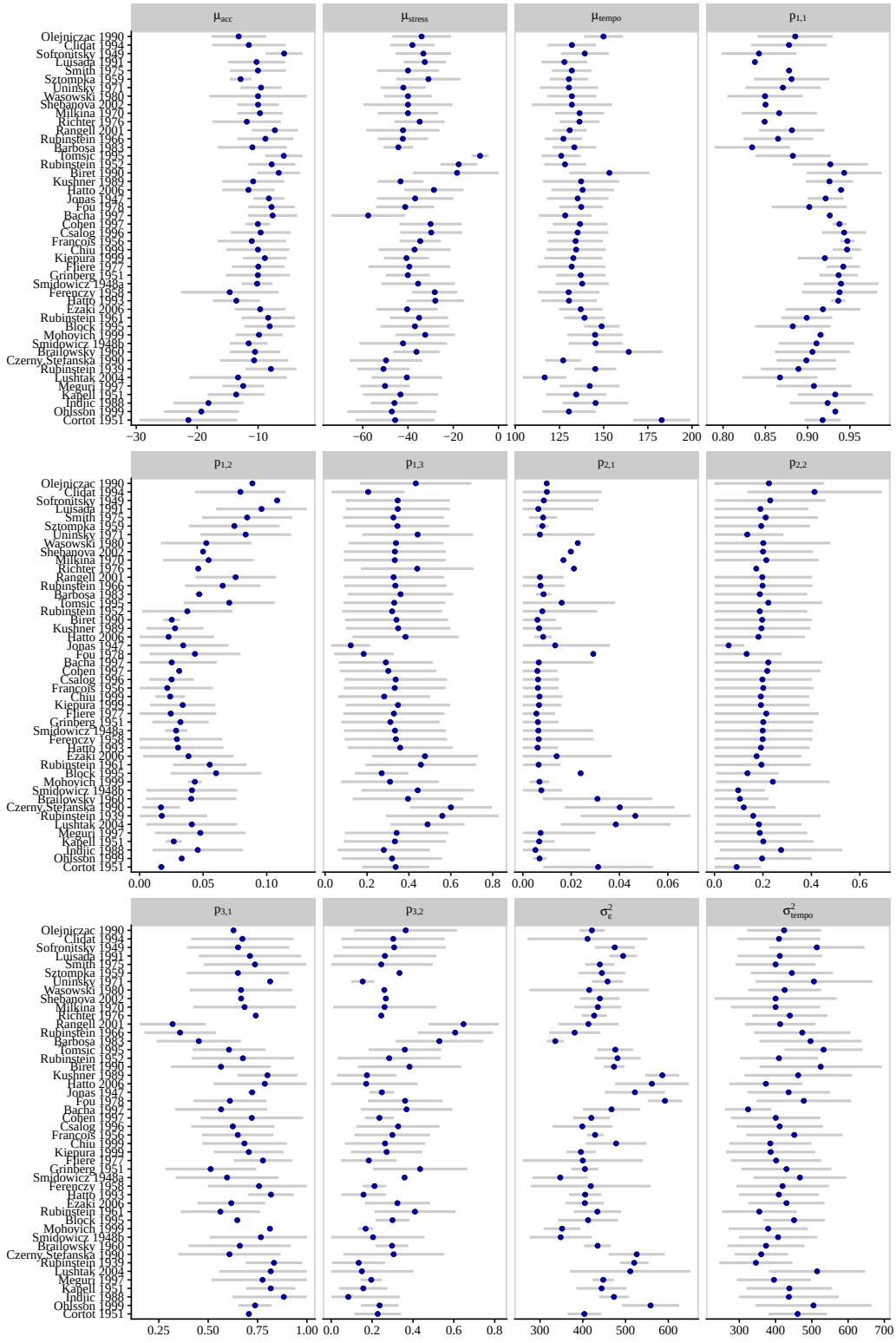

FIG SM-1. *Confidence intervals for all parameters based on the observed Fisher information.*

Recordings in the fourth group (Figure SM-4, $\oplus$ in Figure 10) are those in the upper right of Figure 9, from Olejniczac to Richter. These recordings tend to transition quickly between states, especially constant tempo and slowing down accompanied by frequent transitory emphases. The probability of remaining in state 1 is the lowest while the probability of entering state 2 from state 1 is the highest. The acceleration state is rarely visited. Four of the most similar performances are in this group along with Richter's 1976 recording.

The three performances in group six (Figure SM-5, $\bullet$ in Figure 10) are actually quite like others, but with small exceptions. Biret's 1990 performance is very much like those in group 1, but with a much larger contrast between tempos in the A and B sections. The recording by Rubinstein in 1952 is similar, though with a faster A section that has less contrast with the B section. Tomsic's 1995 performance is actually most similar to those in group three ($\ast$), but played much faster and with a large σ^2_ϵ.

The remaining performances are displayed in Figures SM-6–SM-9, with the exception of Cortot's performance in the manuscript.

6. Distribution over states. To examine the stability of the Algorithm 1, we examined all the potential paths for Richter's 1976 recording. Here, we saved the most likely 10,000 paths and their weights (rather than only the most likely path). Figure SM-10 shows the marginal (posterior) probability of being in a particular state for each note. While the paper uses the most likely *path*, this figure is marginal in the sense that a particular note/state combination will have high probability when many paths visited that note/state. But, the most likely path may not have used that same note/state combination. Nonetheless, there appears to be consensus for many of the notes. The most obviously difficult notes are those near measures 10 and 50. In both cases, the most likely path (Figure 1 in the main text) used the stress state, which exceeds 50% posterior probability here.

7. Multiplicative tempo changes. While an additive state space model is relatively easy to understand, some music theorists (Mead, 2007, for example) have argued that musicians make multiplicative tempo adjustments. That is, it is the ratio between the tempo of the current note and that of the previous note rather than their difference that is important. Such

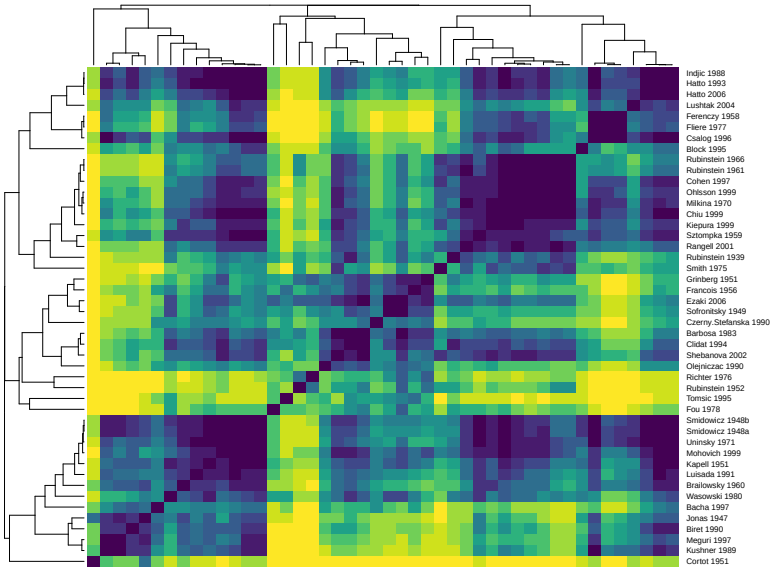

FIG SM-2. *This figure presents a heatmap and hierarchical clustering based only on the note-by-note onset timings for each of the 46 recordings.*

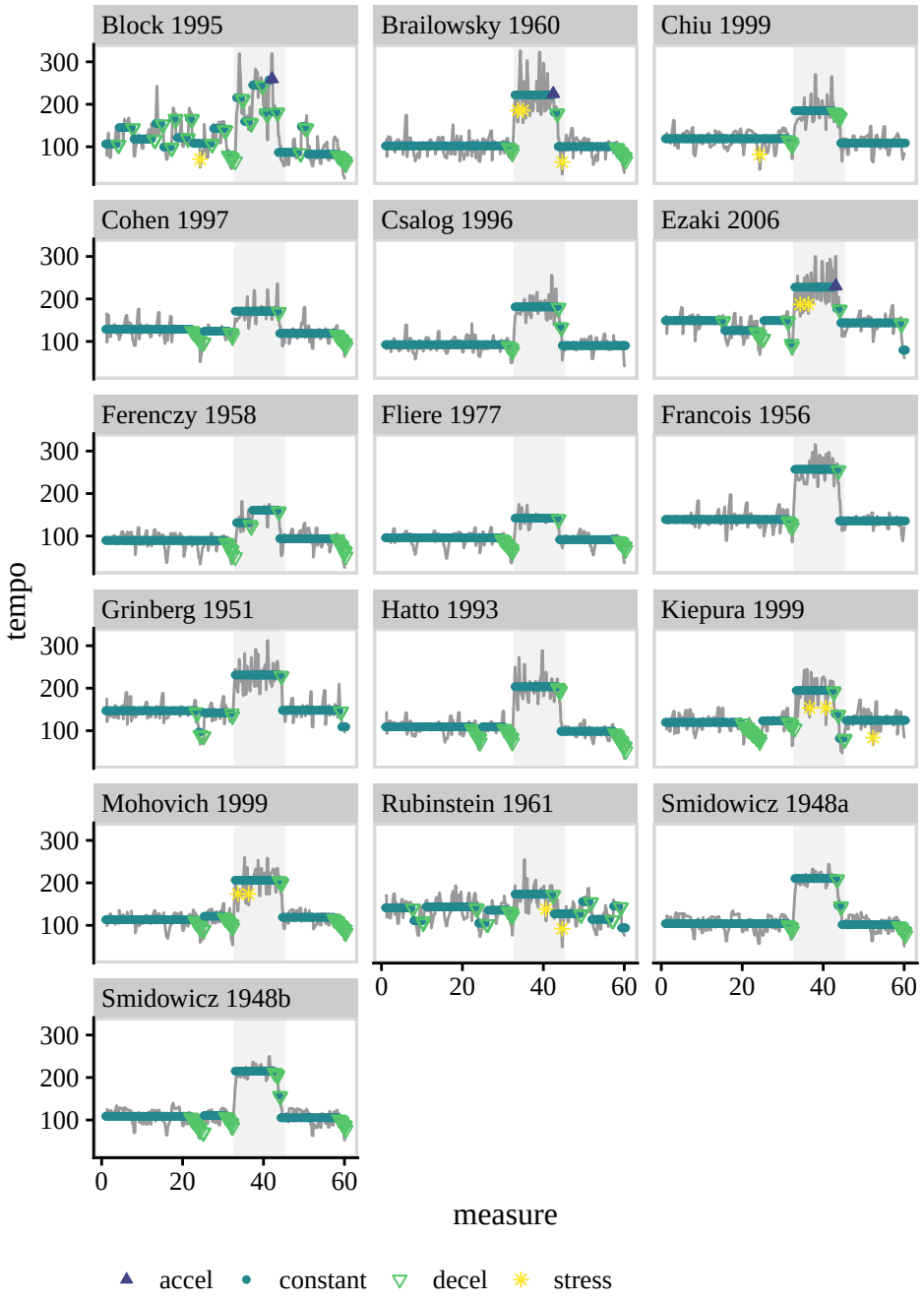

FIG SM-3. *Performances in the first group*

a conception is fundamental to musical notation (quarter notes, eighth notes, etc) and frequently used to specify tempo changes within a piece of music, such as with 𝅗𝅥 = ♩ to indicate that the next section should be played at half the previous tempo.

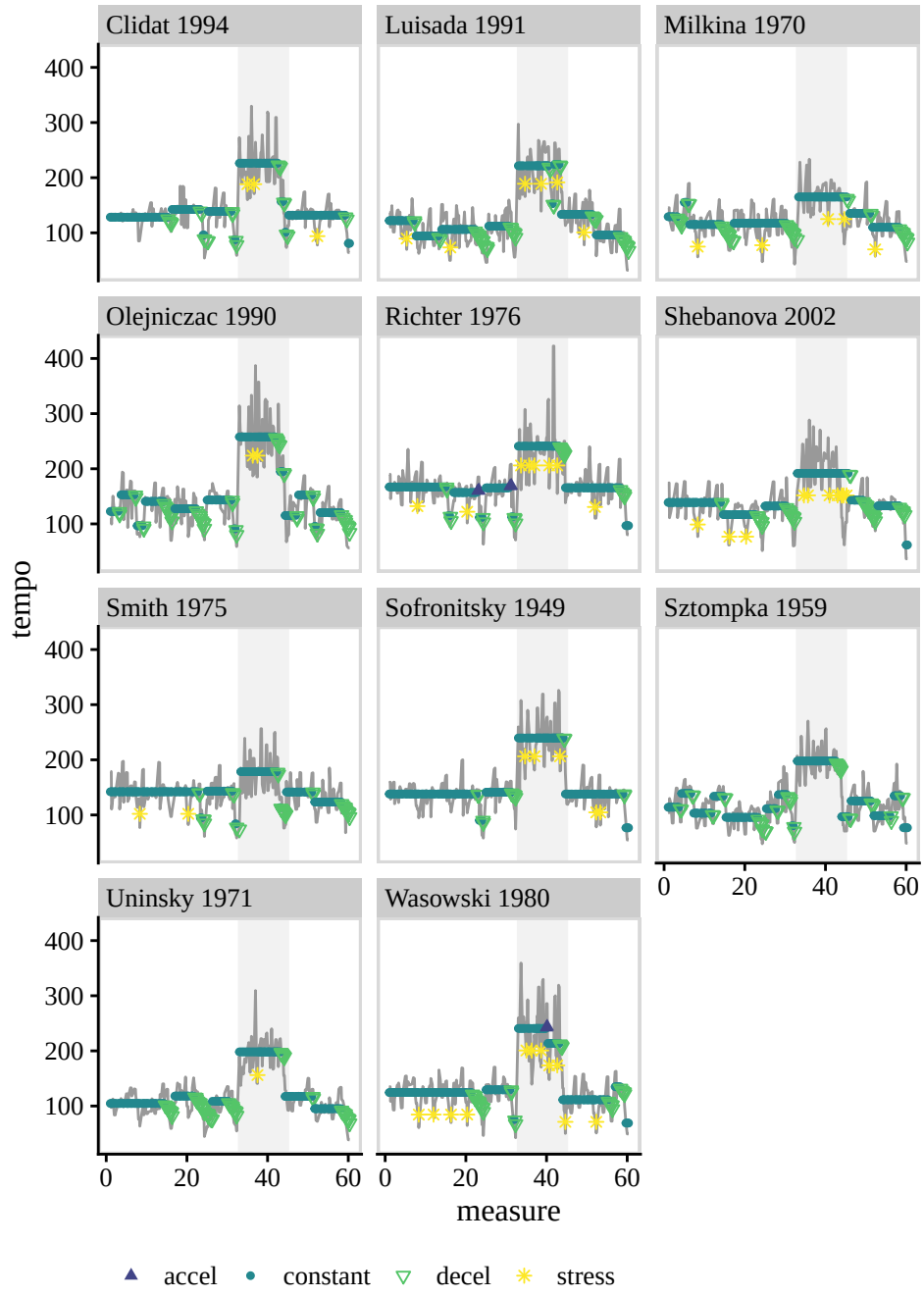

FIG SM-4. *Performances in the fourth group.*

Rather than the linear switching model described in Section 2, we examined the following multiplicative version:

$$
\begin{aligned}
x_1 &\sim \text{lognormal}(x_0,\ P_0), \\
\frac{x_{i+1}}{x_i} &= (1-\mu(s_i))\eta_i, \qquad \eta_i \sim \text{lognormal}(0,\ Q(s_i)), \\
y_i &= c(s_i) + x_i + \epsilon_i, \qquad \epsilon_i \sim N(0,\ G(s_i)).
\end{aligned}
$$

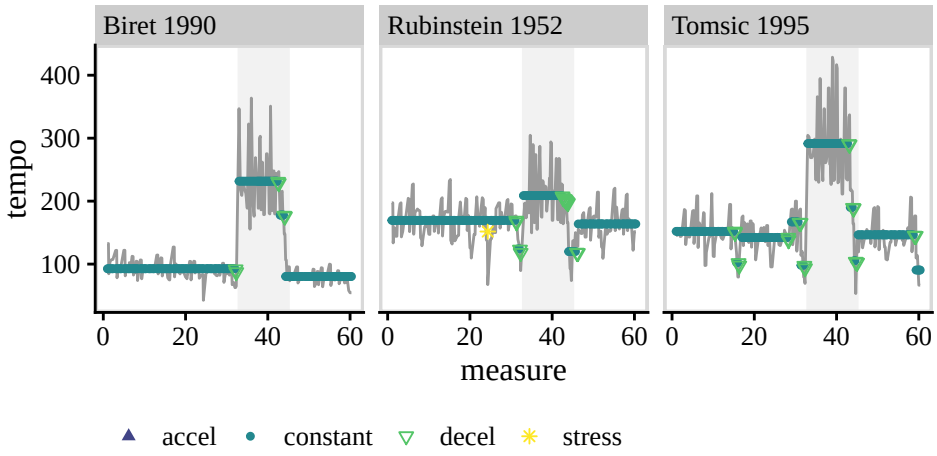

FIG SM-5. *Performances in the sixth group.*

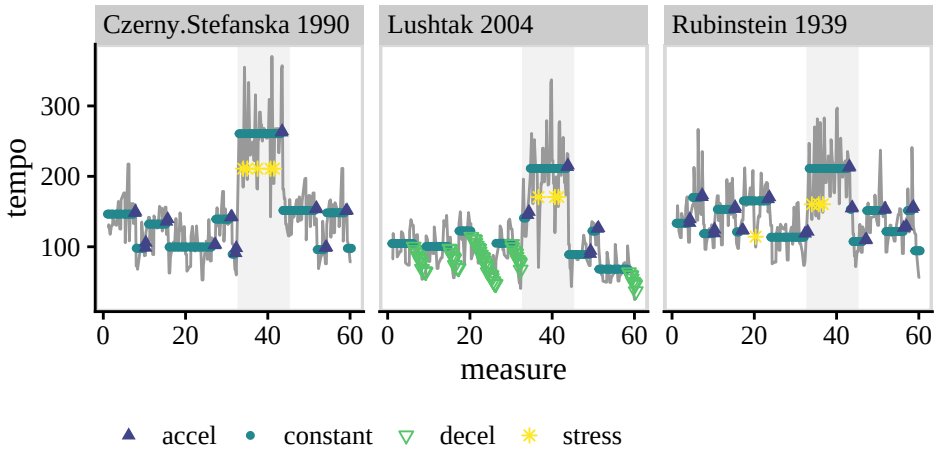

FIG SM-6. *Performances in the second group.*

Here, $\mu(s_i)$ is 0 in the constant tempo or emphasis states and controls the magnitude of acceleration or deceleration in the other states. To complete the model, $G = \sigma_\epsilon^2 + \sigma_{stress}^2 I(s_i = 4)$ while $Q = \sigma_{acc}^2$ in states 2 or 3 and 0 otherwise. To make this model easier to compute, we transform by examining the log of the transition equation and exponentiating the hidden continuous state in the measurement equation. Likelihood evaluation is then performed with the extended Kalman filter (EKF). The EKF is essentially the Kalman filter applied to the first-order Taylor series expansion of any non-linear components around our current predictions of them. For this model, we have

$$\begin{aligned} \log(x_1) &\sim N(x_0,\ P_0), \\ \log(x_{i+1}) = \log(x_i) + \log(1 - \mu(s_i)) + \log(\eta_i), &\quad \log(\eta_i) \sim N(0,\ Q(s_i)), \\ y_i = c(s_i) + \exp(\chi_i) + \exp(\chi_i) x_i + \epsilon_i, &\quad \epsilon_i \sim N(0,\ G(s_i)), \end{aligned}$$

where χ_i is the estimate of $E\left[x_i \mid y_1, \ldots y_{i-1}\right]$.

We estimated this same model on the entire dataset and performed principal component analysis on the resulting parameters (see also Section 3.3 of the manuscript). Figure SM-11 shows the inferred performance decisions for Richter's 1976 performance from both the linear and multiplicative models. Both seem to fit the data quite well. There are three slight

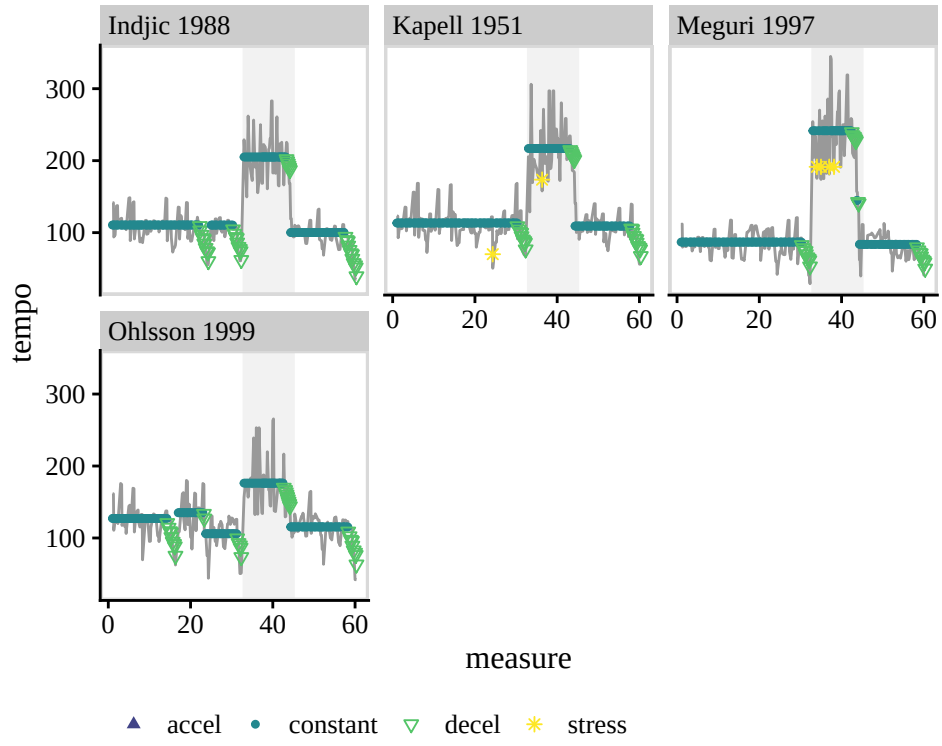

FIG SM-7. *Performances in the third cluster.*

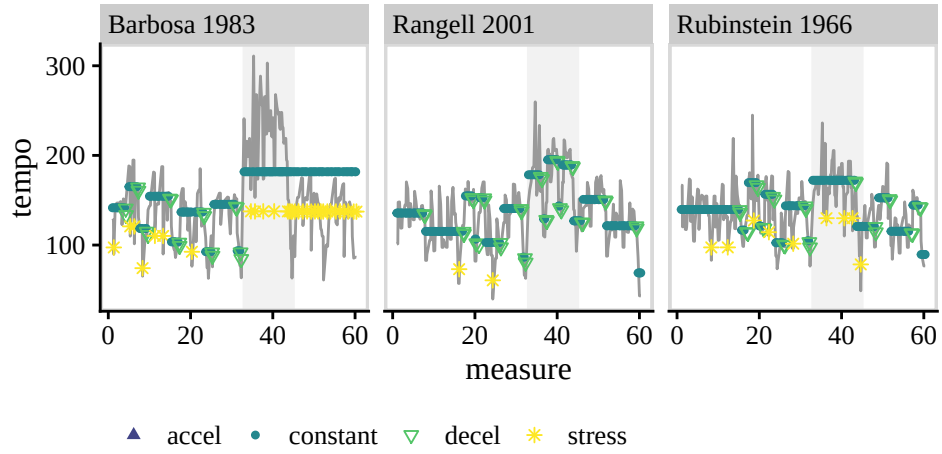

FIG SM-8. *Performances in the fifth group.*

differences between the inferences. First, in the B section, the multiplicative model uses the acceleration state more than the additive model does. Second, the multiplicative model avoids the stress state, and this behavior is reflected in a higher probability of remaining in the constant tempo state. Third, the perioods of slowing down at the ends of phrases are better explained by the multiplicative model.

The percent of variance explained by the first two principal components is 99%. The first factor loads completely on σ_ϵ^2 while the second loads on μ_{stress}. So, while this model can explain individual performances quite well, it is much less able to provide musically mean-

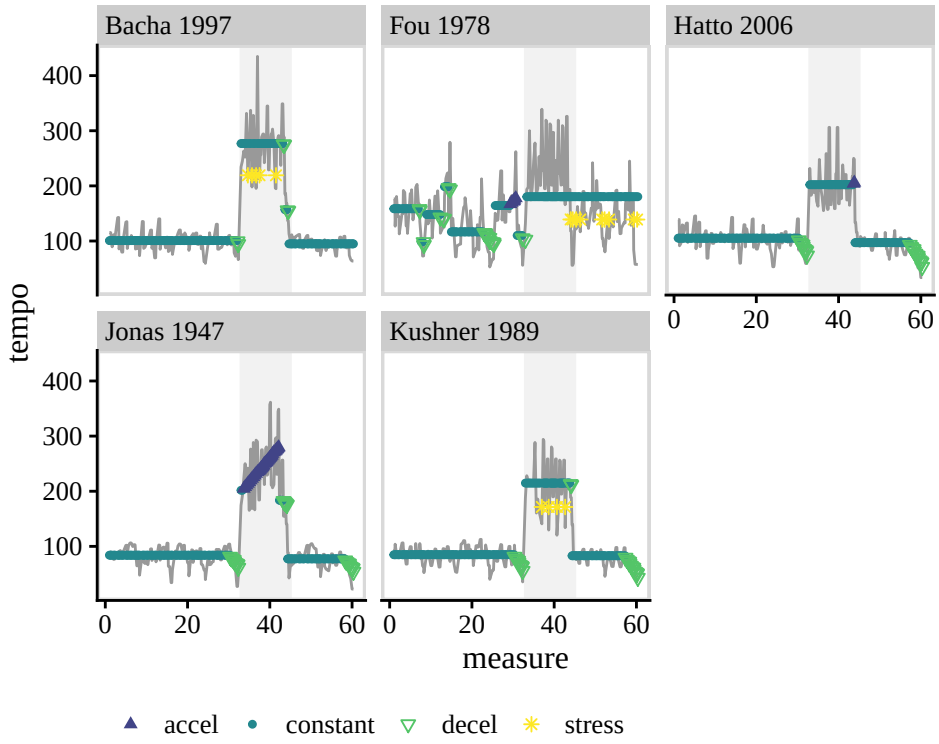

FIG SM-9. *Performances in the seventh group.*

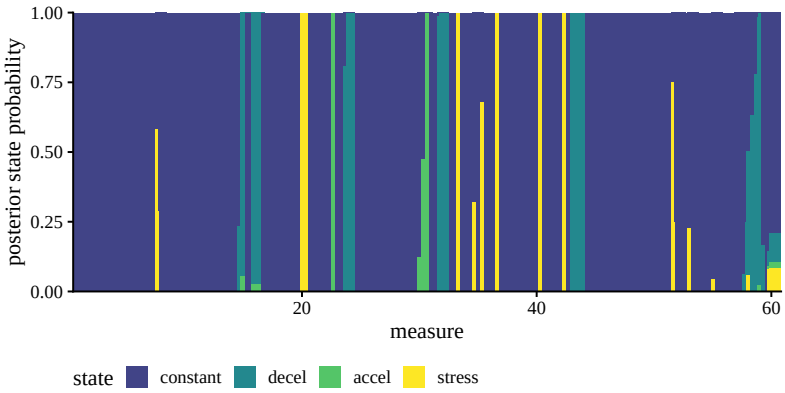

FIG SM-10. *Distribution over potentitial states for Richter's 1976 recording.*

ingful distinctions between performances. While the interpretation for Richter's performance seems quite reasonable under this model, other performances are much less reasonable.

8. Alternative prior distributions. Following the recommendations of an anonymous referee, we reestimated the model under some alternative prior distributions. These distributions are shown in Table 2. Returning to Richter's recording, we used inverse gamma and uniform distributions for the variance parameters to allow heavier tails. We also looked at uniform distributions on the transition probabilities and a prior which requires the observation variance, σ_ϵ^2 to be smaller.

Apart from the "smaller observation variance" setting, these different specifications do not have a dramatic effect: the fit to the data remains similar both quantitatively (as mea-

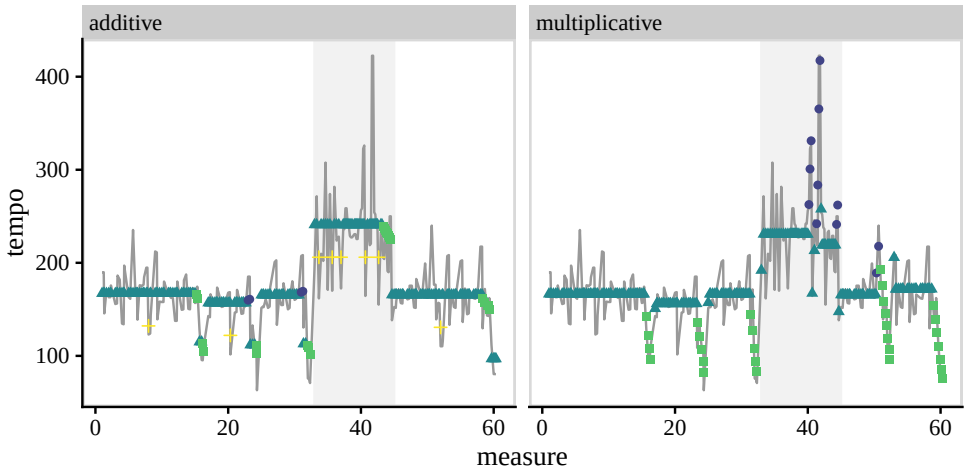

FIG SM-11. *Additive and multiplicative models for Richter's 1976 performance. The multiplicative model fits quite well, but is less likely to visit the stress state.*

sured by RMSE and negative loglikelihood, see Table 3) and qualitatively (as determined by examining the inferred performance in Figure SM-13).

The prior modes are important for some parameters to avoid non-identifiability, and occasionally, as described in the manuscript, to enforce more musically meaningful switching behaviors. On the other hand, the prior tail shape is not particularly important here because we're estimating posterior modes rather than performing a full Bayesian analysis with accompanying credible intervals.

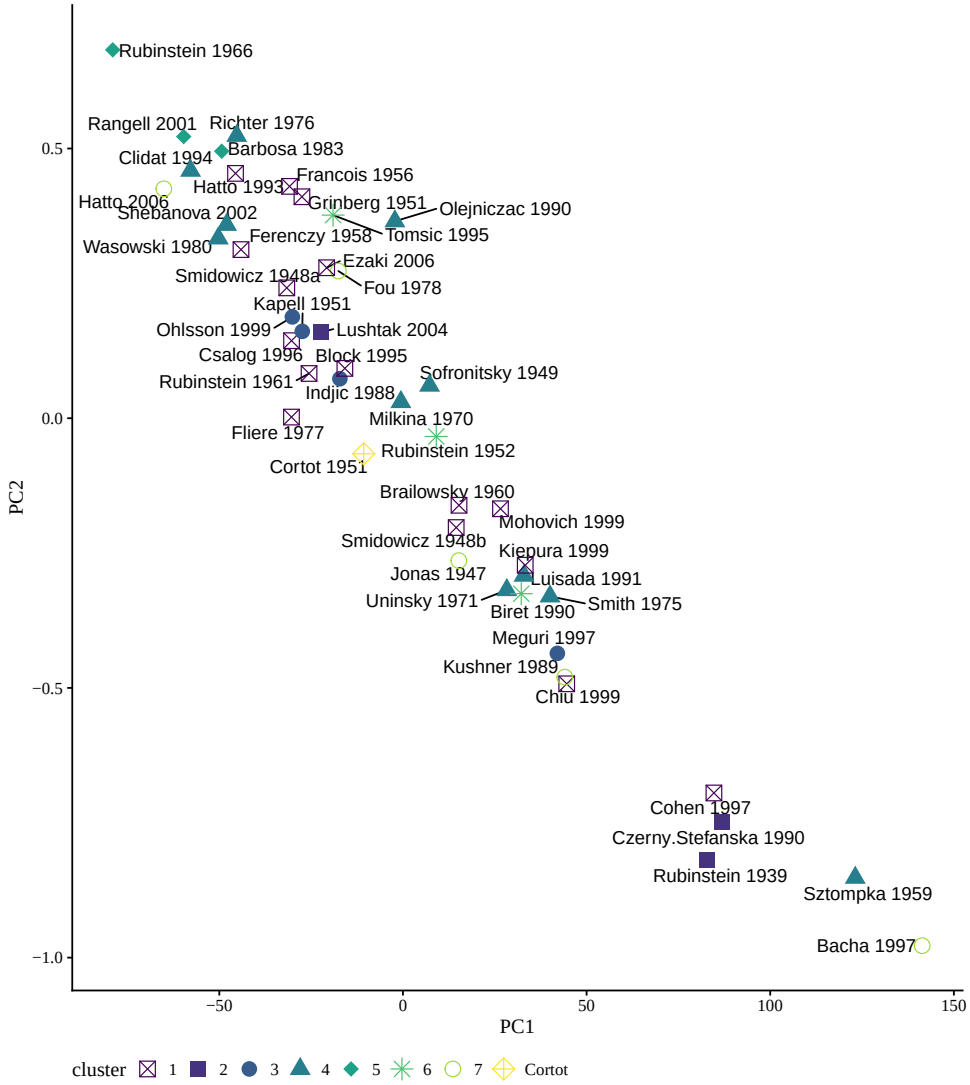

FIG SM-12. *The first two principal components based on parameter estimates from the multiplicative model. The groupings (color and point type) are the same as those from the linear model.*

TABLE 2
Informative prior distributions for the music model

Parameter		Original	Inverse Gamma
σ^2_ϵ	$\sim$	Gamma$(40,\ 10)$	IG$(42, 16400)$
μ_{tempo}	$\sim$	Gamma$\left(\frac{\overline{Y}^2}{100},\ \frac{100}{\overline{Y}}\right)$	IG$\left(\frac{\overline{Y}^2}{100}+2, \overline{Y}(\frac{\overline{Y}^2}{100}+1)\right)$
$-\mu_{\text{acc}}$	$\sim$	Gamma$(15,\ 2/3)$	IG(17, 160)
$-\mu_{\text{stress}}$	$\sim$	Gamma$(20,\ 2)$	IG(22, 840
σ^2_{tempo}	$\sim$	Gamma$(40,\ 10)$	IG$(42, 16400)$
σ^2_{acc}	$=$	1	1
σ^2_{stress}	$=$	1	1
$p_{1,\cdot}$	$\sim$	Dirichlet$(85,\ 5,\ 2,\ 8)$	Dirichlet$(85,\ 5,\ 2,\ 8)$
$p_{2,\cdot}$	$\sim$	Dirichlet$(4,\ 10,\ 1,\ 0)$	Dirichlet$(4,\ 10,\ 1,\ 0)$
$p_{3,\cdot}$	$\sim$	Dirichlet$(5,\ 3,\ 7,\ 0)$	Dirichlet$(5,\ 3,\ 7,\ 0)$

Parameter		Smaller σ^2_ϵ	Uniform Variances	Uniform Probabilities
σ^2_ϵ	$\sim$	Gamma$(20,\ 10)$	1	Gamma$(20,\ 10)$
μ_{tempo}	$\sim$	Gamma$\left(\frac{\overline{Y}^2}{100},\ \frac{100}{\overline{Y}}\right)$	Gamma$\left(\frac{\overline{Y}^2}{100},\ \frac{100}{\overline{Y}}\right)$	Gamma$\left(\frac{\overline{Y}^2}{100},\ \frac{100}{\overline{Y}}\right)$
$-\mu_{\text{acc}}$	$\sim$	Gamma$(15,\ 2/3)$	Gamma$(15,\ 2/3)$	Gamma$(15,\ 2/3)$
$-\mu_{\text{stress}}$	$\sim$	Gamma$(20,\ 1)$	Gamma$(20,\ 2)$	Gamma$(20,\ 2)$
σ^2_{tempo}	$\sim$	Gamma$(40,\ 10)$	1	Gamma$(40,\ 10)$
σ^2_{acc}	$=$	1	1	1
σ^2_{stress}	$=$	1	1	1
$p_{1,\cdot}$	$\sim$	Dirichlet$(85,\ 5,\ 2,\ 8)$	Dirichlet$(85,\ 5,\ 2,\ 8)$	1
$p_{2,\cdot}$	$\sim$	Dirichlet$(4,\ 10,\ 1,\ 0)$	Dirichlet$(4,\ 10,\ 1,\ 0)$	1
$p_{3,\cdot}$	$\sim$	Dirichlet$(5,\ 3,\ 7,\ 0)$	Dirichlet$(5,\ 3,\ 7,\ 0)$	1

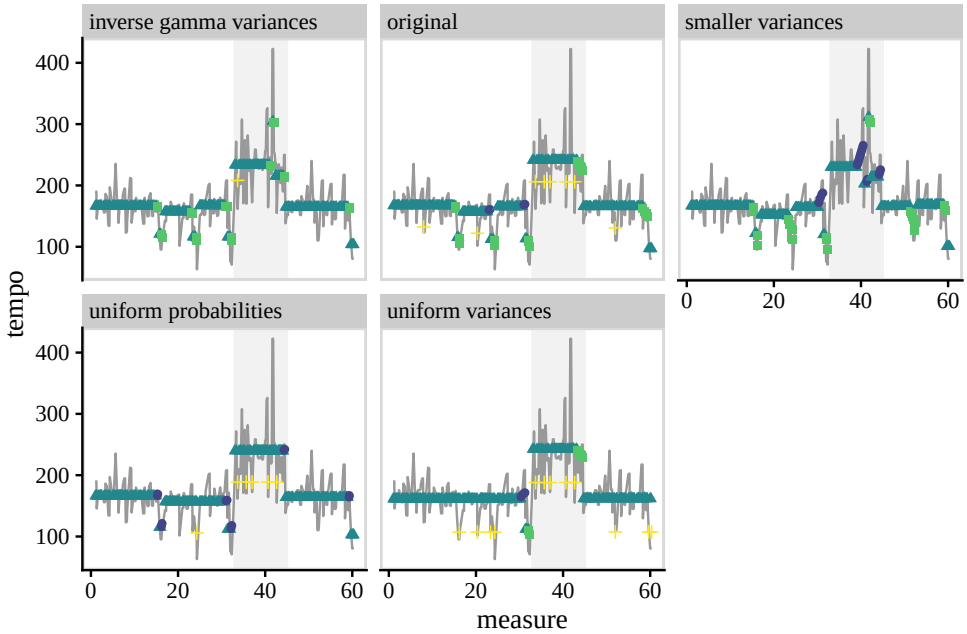

FIG SM-13. *Inferred state sequence for Richter's 1976 recording under alternative prior specifications.*

TABLE 3
For each of the different prior distributions, we report the MSE between the estimated performer intentions from the model and the true performance as well as the negative log likelihood.

prior	rmse	negative log likelihood	σ_ϵ
inverse gamma variances	28.57	-5.20	20.66
original	29.31	-5.10	19.24
smaller variances	26.77	-5.10	20.01
uniform probabilities	30.54	-5.06	21.27
uniform variances	31.05	-5.15	23.56

TABLE 4
For each of the different prior distributions, we report the estimated parameter values.

	original	smaller variances	inverse gamma variances	uniform variances	uniform probabilities
σ^2_ϵ	426.70	370.37	400.36	452.49	555.29
μ_{tempo}	136.33	167.65	166.55	133.34	135.69
μ_{acc}	-11.84	-23.55	-6.44	-10.02	-7.74
μ_{stress}	-34.82	-16.76	-24.75	-54.16	-50.89
σ^2_{tempo}	439.38	451.58	400.57	406.50	320.17
p_{11}	0.85	0.94	0.91	0.90	0.93
p_{12}	0.05	0.03	0.05	0.02	0.01
p_{22}	0.74	0.46	0.52	0.68	0.85
p_{31}	0.44	0.22	0.30	0.29	0.74
p_{13}	0.02	0.01	0.01	0.01	0.03
p_{21}	0.25	0.47	0.45	0.26	0.08
p_{32}	0.17	0.04	0.20	0.13	0.15

REFERENCES

ANDERSON, B. D., AND MOORE, J. B. (1979), *Optimal filtering*, Prentice-Hall, Englewood Cliffs, NJ.

MEAD, A. (2007), "On tempo relations," *Perspectives of New Music*, **45**(1), 64–108.

RAUCH, H. E., STRIEBEL, C., AND TUNG, F. (1965), "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, **3**(8), 1445–1450.