USING THE DISTRIBUTION DERIVATIVE METHOD TO MODEL ACOUSTIC MUSICAL INSTRUMENT SOUNDS WITH POLYNOMIAL AM-FM SINUSOIDS

Marcelo Caetano

Department of Music
UC San Diego
La Jolla, CA, USA
mcaetano@ucsd.edu

ABSTRACT

The oscillatory modes of musical instrument sounds are commonly modeled with time-varying sinusoids. Several estimation methods model quasi-stationary oscillations accurately, yielding a high-quality representation. However, nonstationary oscillations such as attack transients are still very challenging to model accurately. In this work, we propose to model musical instrument sounds with polynomial modulation sinusoids (PMS) estimated with the distribution derivative method (DDM). DDM gives accurate parameter estimations for PMS with arbitrary order, allowing great flexibility in modeling temporal changes inside analysis frames as amplitude and frequency modulations. We used 39 musical instrument sounds to compare DDM objectively against the standard (SM+) and an adaptive sinusoidal model (eaQHM) using time and frequency error measures. We showed that DDM captures more oscillatory energy than SM+ or eaQHM by modeling PMS more accurately. A MUSHRA listening test with 18 selected sounds confirmed that DDM has higher perceptual quality than both SM+ and eaQHM and that DDM is almost perceptually indistinguishable from the original sounds.

1. INTRODUCTION

Acoustic musical instruments are particularly challenging to model accurately due to the very characteristics that make them musically interesting, such as attack transients, inharmonicity, modal coupling, among many others. Sinusoidal models [1, 2, 3, 4, 5, 6, 7, 8] stand out among the representations of musical instrument sounds due to their fidelity and flexibility. The sinusoidal model assumes that, inside a time frame, a sound $y\left(t\right)$ can be decomposed as

$$y\left(t\right) = s\left(t\right) + r\left(t\right), \tag{1}$$

where the sinusoidal component $s\left(t\right)$ captures the oscillatory modes with time-varying sinusoids and the modeling residual $r\left(t\right)$ contains everything that was not captured by the sinusoids, such as instrumental and additive background noise [2, 9, 10].

The quality of sinusoidal models depends not only on the accuracy of estimation of the sinusoids, but also on how well the model assumptions fit the characteristics of the sound. Traditionally, sinusoidal models represent musical instrument sounds as a sum of piecewise *quasi-stationary* sinusoids (QSS) under the assumption that the partials are relatively stable inside each frame [1, 2]. The estimation of QSS has a long history in the literature [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23], with the most common techniques relying on frequency bins around spectral peaks, such as parabolic interpolation [1, 2, 20, 21, 22, 23]. However, nonstationary oscillations that vary inside a frame, such as attack transients, are poorly modeled by QSS [24, 8].

Several authors have proposed to use *nonstationary* sinusoids to more accurately capture modulations in amplitude and frequency within an analysis frame [3, 4, 5, 25, 26, 6, 8, 27]. In these models, the sinusoidal component $s\left(t\right)$ is further modeled with nonstationary sinusoids as

$$s\left(t\right) = \sum_{k=1}^{K} x_k\left(t\right) = \sum_{k=1}^{K} \exp\left\{A_k\left(t\right) + j\,\Phi_k\left(t\right)\right\}, \tag{2}$$

where k is the index of each *partial* $x_k\left(t\right)$, K is the maximum number of partials, $A_k\left(t\right)$ is the instantaneous log-amplitude and $\Phi_k\left(t\right)$ is the instantaneous phase of $x_k\left(t\right)$. Each partial $x_k\left(t\right)$ in eq. (2) is called an AM-FM sinusoid [28, 26, 27] because its amplitude is modulated by $A_k\left(t\right)$ and its frequency by the first time derivative of $\Phi_k\left(t\right)$.

Using the same parametric model to represent both the AM and the FM components, we write each partial $x_k\left(t\right)$ (dropping the index k for simplicity) as a *polynomial modulation* sinusoid (PMS) [29, 30, 31, 32] as follows

$$x\left(t\right) = \exp\left\{\sum_{q=0}^{Q} \alpha_q t^q\right\} = \exp\left\{\alpha_0 + \sum_{q=1}^{Q} \alpha_q t^q\right\}, \tag{3}$$

where $\alpha_q = a_q + j\,b_q \in \mathbb{C}$ and Q is the model order. The PMS in eq. (3) are a natural extension of QSS that can represent a broad class of signals that are typically classified according to the order Q and the phase. For example, *linear* phase for $Q = 1$ (i.e., stationary sinusoid when $a_1 = 0$ and exponentially damped sinusoid (EDS) [33] otherwise), *quadratic* phase for $Q = 2$ (i.e., chirp when $a_1 = a_2 = 0$) [28, 5, 34], and *cubic* phase for $Q = 3$ [1, 32]. In the literature, PMS up to $Q = 5$ [35] have been proposed, but the estimation accuracy was not high enough to lead to an increase in modeling quality worth the additional estimation effort. One of the contributions of this work is the use of DDM to estimate PMS for musical instrument sounds because increasing the model order Q with DDM is effortless (see Sec. 2).

Estimation of the parameters of PMS is a particularly challenging problem [36, 29, 37, 31, 38, 32, 27]. As such, many authors have proposed estimation methods for specific cases, such as EDS [33] and quadratic phase (i.e., *chirps*) [4, 28, 39, 34]. The reassignment method [40, 41] leads to accurate estimation and high

quality representation of musical instrument sounds [41], and the method of derivatives [39] has also been used to estimate *nonstationary* sinusoids for musical instrument modeling [32, 7]. Fullband modeling with adaptive sinusoids achieved state-of-the-art performance for musical instrument sounds [8]. The extended adaptive Quasi-Harmonic Model (eaQHM) [42] iteratively fits a harmonic template to the full frequency range of the spectrum. Thus, adaptation yields a high-quality representation of nonstationary sinusoids but it also models noisy oscillations, especially at high frequencies.

Betser developed the distribution derivative method (DDM) [29], capable of estimating the parameters of PMS of arbitrary order Q with high accuracy [29, 37, 38]. Later, several authors have shown independently that, under certain conditions, DDM is conceptually equivalent [29, 37, 31] to the reassignment method [40, 41] and to the method of derivatives [39]. However, these works either compare theoretical results or the estimation accuracy of synthetic sinusoids against the theoretical Cramér-Rao bound (CRB) [29, 37, 38]. To the best of our knowledge, no previous work has used DDM to model sounds. The main contributions of this work are the description of the implementation of an AM-FM sinusoidal model that uses DDM to estimate PMS for musical instrument sounds and the comparison with other sinusoidal models.

Sec. 2 presents a brief overview of DDM estimation of PMS with arbitrary order with the short time Fourier transform (STFT). Next, Sec. 3 describes the objective and perceptual evaluation protocol to compare DDM modeling of musical instrument sounds with the literature. Then, Sec. 4 presents the results of the comparison followed by a discussion. Finally, Sec. 5 presents the conclusions and future work.

2. DISTRIBUTION DERIVATIVE METHOD

The foundations of DDM estimation lie in the theory of distributions in mathematical analysis, which generalizes the concept of functions to include objects such as the Dirac delta [43, 44] with the aid of the test function ψ with compact support U_ψ. Distributions x are interpreted as acting on the test function ψ via integration over the support U_ψ. Notably, the theory of distributions extends the notion of derivative of x, which lies at the core of DDM. Specifically, we want to analyze the signal $x\,(t)$ with the test function $\psi\,(t)$, which is zero outside U_ψ and differentiable at least once on U_ψ. For such, we use the inner product

$$\langle x, \psi \rangle = \int_{-\infty}^{\infty} x\,(t)\,\bar{\psi}\,(t)\,\mathrm{d}t = \int_{U_\psi} x\,(t)\,\bar{\psi}\,(t)\,\mathrm{d}t \tag{4}$$

where $\bar{\psi}$ is the complex conjugate of ψ. The *derivative* of x with respect to (w.r.t.) its argument t, which we denote $\dot{x}$, is obtained with the aid of $f\,(t) = x\,(t)\,\bar{\psi}\,(t)$. The derivative of f w.r.t. t is then

$$\dot{f}\,(t) = \dot{x}\,(t)\,\bar{\psi}\,(t) + x\,(t)\,\dot{\bar{\psi}}\,(t)\,. \tag{5}$$

Integrating eq. (5) w.r.t t over the support U_ψ yields

$$\int_{U_\psi} \dot{f}\,(t)\,\mathrm{d}t = \int_{U_\psi} \dot{x}\,(t)\,\bar{\psi}\,(t)\,\mathrm{d}t + \int_{U_\psi} x\,(t)\,\dot{\bar{\psi}}\,(t)\,\mathrm{d}t. \tag{6}$$

Here we note that $\int_{U_\psi} \dot{f}\,(t)\,\mathrm{d}t = 0$ because ψ is strictly nonnegative and vanishes at the borders of U_ψ. Then, the inner product in eq. (4) gives

$$\langle \dot{x}, \psi \rangle = -\langle x, \dot{\psi} \rangle, \tag{7}$$

which allows implicitly taking the derivative of the unknown signal x by differentiating ψ instead. DDM [29] applies eq. (7) to the PMS signal model of eq. (3) to estimate the parameters α_p. Differentiating eq. (3) gives

$$\dot{x}\,(t) = \sum_{q=1}^{Q} \alpha_q q\, t^{q-1} x\,(t)\,, \tag{8}$$

which only depends on t and q. Replacing eq. (8) in eq. (7) gives

$$\sum_{q=1}^{Q} \alpha_q \langle q\, t^{q-1} x, \psi \rangle = -\langle x, \dot{\psi} \rangle, \tag{9}$$

which is the DDM estimation equation for the PMS in eq. (3). Note that eq. (9) is linear w.r.t the PMS parameters α_q. Sec. 2.1 derives $\langle x, \dot{\psi} \rangle$ with the *windowed* Fourier transform. Sec. 2.2 explains how to use the discrete Fourier transform (DFT) to estimate α_q as the linear coefficients of a system of equations given by eq. (9) because of the widespread availability of fast Fourier transform (FFT) implementations and effortless adaptability of the method for the short-time Fourier transform (STFT). Finally, Sec. 2.3 presents the estimation of the remaining parameter α_0.

2.1. DDM Estimation with the Windowed Fourier Transform

For the *windowed* Fourier transform, $\psi\,(t) = w\,(t)\,\mathrm{e}^{j\omega t}$, where $w\,(t)$ is a tapering window with compact support that is differentiable at least once. Then,

$$\dot{\bar{\psi}}\,(t) = [\dot{w}\,(t) - j\omega w\,(t)]\,\mathrm{e}^{-j\omega t}. \tag{10}$$

Replacing eq. (10) on the right-hand side of eq. (7), we get

$$\langle x, \dot{\psi} \rangle = \langle x, \tau \rangle - j\omega \langle x, \psi \rangle, \tag{11}$$

where we use the shorthand $\tau(t) = \dot{w}(t)e^{j\omega t}$ to define $\langle x, \tau \rangle$ as

$$\langle x, \tau \rangle = \langle x, \dot{w}\, e^{j\omega t} \rangle. \tag{12}$$

Equation (11) shows that $\langle x, \dot{\psi} \rangle$ only requires computing two windowed Fourier transforms of the signal $x\,(t)$, namely $\langle x, \psi \rangle$ and $\langle x, \tau \rangle$ in eq. (12), which simply uses the first time derivative of the window $w\,(t)$. Note that $\dot{w}\,(t)$ can be computed analytically for most commonly used windows (except the rectangular window, which has discontinuities, and the slepian window, which does not have an explicit analytical expression), making the computation of eq. 12 very efficient. Also note that ω in eq. (11) is the frequency from the Fourier kernel $\psi\,(t)$ from the derivative in eq. (10).

2.2. DDM estimation of PMS with the DFT

Combined, eq. (9), eq. (11), and eq. (12) allow estimating the parameters α_q of PMS with the *windowed* DFT as

$$\begin{bmatrix} \langle x, \psi_k \rangle & \langle 2tx, \psi_k \rangle & \cdots & \langle Qt^{Q-1}x, \psi_k \rangle \\ \langle x, \psi_{k+1} \rangle & \langle 2tx, \psi_{k+1} \rangle & \cdots & \langle Qt^{Q-1}x, \psi_{k+1} \rangle \\ \vdots & \vdots & & \vdots \\ \langle x, \psi_{k+R} \rangle & \langle 2tx, \psi_{k+R} \rangle & \cdots & \langle Qt^{Q-1}x, \psi_{k+R} \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_Q \end{bmatrix} = - \begin{bmatrix} \langle x, \dot{\psi}_k \rangle \\ \langle x, \dot{\psi}_{k+1} \rangle \\ \vdots \\ \langle x, \dot{\psi}_{k+R} \rangle \end{bmatrix}, \tag{13}$$

where $\psi_k = w\,(n)\,\mathrm{e}^{j\,\omega_k n}$ corresponds to a bin k of the DFT and R is the total number of bins used in the estimation. See [29] for further details on the discretization. Equation (13) can be solved uniquely for $R \geq Q$, and larger R reduces estimation bias when

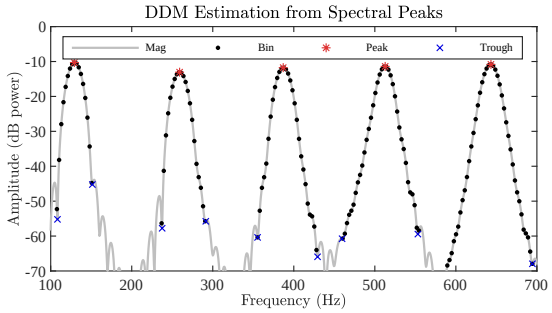


Figure 1: Illustration of DDM estimation from spectral peaks. The spectral peaks ($*$) and troughs ($\times$) are respectively local maxima and minima of the magnitude spectrum (solid line). The spectral bins k in eqs. (13) and (16) are shown as $\bullet$.

ψ_k are taken from the main lobe, where the spectral energy is concentrated [37, 38]. Thus, zero padding increases estimation accuracy at a relatively low computational cost with the FFT. Fig. 1 illustrates the range of k used in the estimation of five PMS. Note that R can be different for each spectral peak depending of its shape, typically requiring independent estimation per peak. Sec. 2.4 describes selection of the spectral peaks and bins used in eq. (13).

2.3. Estimation of α_0

The parameter $\alpha_0 = a_0 + j\, b_0$ is not estimated by eq. (13) because of the derivative in eq. (8). Note that a_0 is the constant amplitude and b_0 is the initial phase of the PMS. Once α_q, $q \in [1, \cdots, Q]$ have been estimated, α_0 can be estimated by writing eq. (3) as

$$x\,(t) = \exp\{\alpha_0\}\,\bar{x}\,(t)\,, \tag{14}$$

with $\bar{x}\,(t) = \exp\left\{\sum_{q=1}^{Q} \alpha_q t^q\right\}$ synthesized with α_q estimated with eq. (13). The inner product from eq. (4) in eq. (14) gives

$$\langle \bar{x}, \psi \rangle \exp\{\alpha_0\} = \langle x, \psi \rangle, \tag{15}$$

which becomes

$$\begin{bmatrix} \langle \bar{x}, \psi_k \rangle \\ \langle \bar{x}, \psi_{k+1} \rangle \\ \vdots \\ \langle \bar{x}, \psi_{k+R} \rangle \end{bmatrix} \exp\{\alpha_0\} = \begin{bmatrix} \langle x, \psi_k \rangle \\ \langle x, \psi_{k+1} \rangle \\ \vdots \\ \langle x, \psi_{k+R} \rangle \end{bmatrix}, \tag{16}$$

which can be estimated by least squares via the pseudo-inverse with the same frequency bins as eq. (13). Fig. 1 illustrates the frequency bins and Sec. 2.4 explains peak selection.

2.4. Peak Selection

Peak selection [45, 46, 47, 48, 49, 50] is used in DDM to avoid mismodeling side lobes or noise as sinusoids. This work uses a simple peak selection scheme based on constant thresholds applied to magnitude spectrum peaks. We use measures of the *absolute level*, *relative level*, average *peak-to-trough difference* (PTD), and *main lobe bandwidth* B to reject peaks as side lobe or noise. Peaks whose maximum is below the absolute level threshold ρ_a are rejected. Similarly, peaks whose maximum is below ρ_r *relative* to the maximum level of the frame are rejected. The average PTD is measured as the average level difference between the peak's maximum and its immediate trough to the left and to the right (see Fig. 1). Peaks whose average PTD is below $\bar{\varepsilon}$ are rejected. Finally, B is measured as the difference in frequency bins between the right and left troughs (see Fig. 1), and peaks whose B is smaller than a threshold ρ_B are rejected.

3. EVALUATION

This section compares the ability of the sinusoidal models SM+ [1, 2] with power scaling estimation [51], eaQHM [42, 8], and DDM [29], to represent musical instrument sounds. Sec. 3.1 describes the sinusoidal models used in the comparison, Sec. 3.2 describes the musical instrument sounds used, Sec. 3.3 describes the objective evaluation, and finally Sec. 3.4 describes the perceptual evaluation. The companion webpage has further information, including the audio files.

3.1. The Sinusoidal Models

SM+ is used as baseline due to its widespread availability, and eaQHM is considered as the state of the art due to its modeling performance (see [8]). SM+ models spectral peaks as stationary sinusoids [1, 2] whose parameters are estimated with power scaling [51], which has been shown to provide an accurate estimation for stationary sinusoids [8].

All common parameters were set to the same value for all models. A zero-phase, non-causal *Hann* window with $M = 6\,T_0\,f_s$ samples was used with 50% overlap, where T_0 is the fundamental period and $f_s = 44.1$kHz the sampling frequency. The FFT size was $N = 2^4 M$ (zero padding). The maximum number of partials was $K = 100$ and $Q = 3$ for DDM. The peak selection thresholds for SM+ and DDM (see Sec. 2.4) were $\rho_a = -110$ dB, $\rho_r = -80$ dB, $\bar{\varepsilon} = 7$ dB, and $\rho_B = 0.8\,\frac{N}{M} B_{\text{mlw}}$, where B_{mlw} is the main lobe width in frequency bins for critical sampling (i.e., $M = N$), thus $B_{\text{mlw}} = 4$ bins for the Hann window.

Table 1 compares the model order of the sinusoidal models used for the analysis and resynthesis stages. SM+ analysis models spectral peaks as stationary sinusoids, whereas resynthesis relies on cubic phase interpolation across frames [1]. Note that the time-varying nature of QSS comes from the resynthesis step translating slight variations in parameter values across frames as slow temporal variations. Analysis in eaQHM is iterative to adaptively capture temporal variations inside each analysis frame as AM-FM sinusoids (see [42, 8] for further details), and synthesis uses cubic phase interpolation with b-splines [42]. Finally, this work uses overlap-add (OLA) synthesis for DDM after each frame is synthesized with eq. (3). Naturally, resynthesis by polynomial phase interpolation in DDM is also possible, but informal listening tests revealed no clear perceptual difference.

Table 1: Comparison of polynomial model order Q between SM+, eaQHM, and DDM.

	Analysis		**Synthesis**	
	Amplitude	*Phase*	*Amplitude*	*Phase*
SM+	Constant	Linear	Linear	Cubic
eaQHM	Adaptive	Adaptive	Linear	Cubic
DDM	(Log) Cubic	Cubic	(Log) Cubic	Cubic

3.2. Musical Instrument Sounds

Table 2 lists the 39 musical instrument sounds from the Brass, Woodwinds, Strings, and Percussion families, which are grouped by playing technique into *sustained*, *percussive*, and *modulated*. Each group has 13 sounds played *forte* or *fortissimo*. Sustained sounds (*ordinario*) are used as baseline for comparison since these contain relatively stable oscillatory modes that can be captured by QSS. Percussive sounds, including plucked strings, are particularly challenging for sinusoidal models [6] because they typically have sharp attacks with transients. Finally, modulation includes both amplitude and frequency modulations, such as *vibrato*, *sforzando*, *pizzicato*, and *staccato*.

3.3. Objective Evaluation

The objective evaluation compares the RMS log spectral measure (RMS-LSM) [52], the Itakura-Saito divergence (ISD) [52], and the signal-to-residual ratio (SRR) [32, 8]. The error between the original spectrum $Y(\theta)$ and its sinusoidal model $S(\theta)$ is defined as [52]

$$V(\theta) = \ln|Y(\theta)|^2 - \ln|S(\theta)|^2, \tag{17}$$

where θ is a normalized frequency with π as the Nyquist frequency. The set of L_p norms defined by [52]

$$d_p = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}|V(\theta)|^p \,\mathrm{d}\theta\right]^{1/p} \tag{18}$$

gives measures of spectral distance d_p, where $p \in \mathbb{Z}_+$ (positive integers). As p increases, the effect of the larger errors is more heavily weighted than the lower errors. RMS-LSM is a true distance metric [52] that can be expressed in decibels when $p = 2$ in eq. (18) by multiplying by $10/\ln(10)$.

The ISD is defined as [52]

$$D = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\mathrm{e}^{V(\theta)} - 1 - V(\theta)\right]\mathrm{d}\theta, \tag{19}$$

based on a likelihood ratio. Positive errors have greater influence than negative ones [52]. ISD is neither a true distance metric nor linearly related to the RMS-LSM [52], but it is thought to better reflect perceptual differences in the spectra. Both RMS-LSM and ISD are calculated for all time frames and the objective evaluation uses the median value shown in Fig. 2.

Finally, the SRR [8, 32] expressed in dB is

$$SRR = 10\log_{10}\left[\frac{\langle y, y\rangle}{\langle r, r\rangle}\right], \tag{20}$$

where $y(t)$ is the original signal, $r(t)$ is the modeling residual as in eq. (2), and $\langle\cdot\rangle$ is the inner product in eq. (4). Thus, the SRR measures how much energy is left in the residual relative to the total energy in the original signal. The SRR is calculated with the full waveform and higher values indicate lower residual energy, reflecting the ability of the model to capture more oscillatory energy from $y(t)$ with sinusoids. The SRR is a *strict* temporal measure because minute estimation errors even in α_0 will cause $r(t)$ to have more energy, lowering the SRR.

Table 2: Musical instrument sounds from VSL and SOL. The dynamics are *fortissimo* (ff) or *forte* (f). The note is C4 unless indicated otherwise. The playing techniques are *ordinario* (od), *vibrato* (v), *sforzando* (z), *pizzicato* (p), and *staccato* (s).

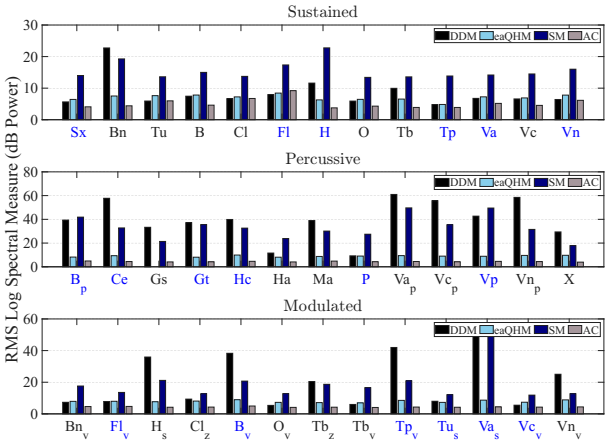
Family	Musical Instrument Sounds
Brass	Bass Tuba (Tu od/s), Trombone (Tb od/v/z), Trumpet C (Tp od/v), French Horn (H od/s)
Woodwinds	Alto Sax (Sx od), Bassoon (Bn od/v), Clarinet B♭ (Cl od/z), Flute (Fl od/v), Oboe (O od/v)
Bowed Strings	Cello (Vc od/p/v), Viola (Va od/p/s), Violin (Vn od/p/v), Contrabass (B od/z/p)
Plucked Strings	Harp (Ha od), Guitar (Gt od), Harpsichord (Hc C♯ od)
Percussion	Glockenspiel (Gs C6 od), Vibraphone (Vp od), Xylophone (X C5 od), Marimba (Ma od), Piano (P od), Celesta (Ce od)

3.4. Perceptual Evaluation

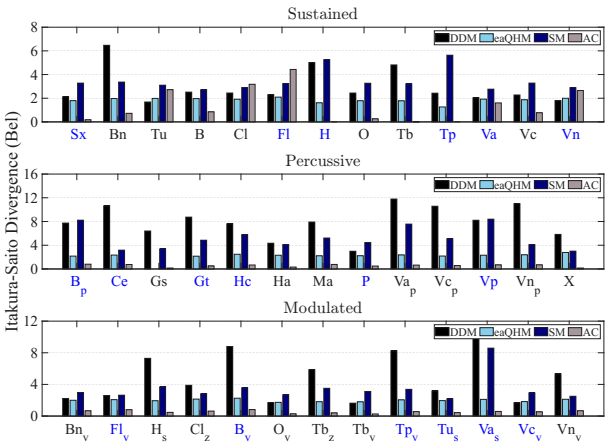
We conducted a listening test to compare the perceptual quality of the different sinusoidal models using the MUSHRA experimental protocol [53]. For each reference sound, participants rated the perceptual quality of each model *relative* to the reference. The reference sound and a degraded version of the reference (dubbed the *anchor*) are also included among the models unbeknownst to the participants. Including the reference allows checking data quality and ensuring that the ratings are assessments of the impairments, whereas the anchor should resemble the typical artifacts of the model being tested. In this work, the anchor is a degraded version of the standard sinusoidal model (only 15 spectral peaks, nearest-neighbor estimation, and resynthesis via phase reconstruction by frequency integration as described in [54]).

Data collection was online and all participants volunteered to take the test. After reading the instructions, participants clicked a button to give informed consent and access the listening test. The test started with an example test page that the participants used to familiarize themselves with the test procedure and also adjust the sound level. The sounds highlighted in Fig. 2 were used in the listening test. In total, 18 sounds were selected, 6 from each playing technique (sustained, percussive, and modulated). Each test page contains the reference at the top and five test sounds being assessed, the three models (SM+, eaQHM, and DDM) plus the reference (REF) and the anchor (AC). Prior to rating the quality of each model, participants were encouraged to listen to the reference before listening to the models. Note that all models are rated simultaneously, so the ratings are expected to reflect relative differences between the models. The order of presentation of the 18 references was randomized for each participant, and so was the order in which the five test sounds appeared.

Participants rated the perceptual quality of the models on a MUSHRA scale from 0 (lowest quality) to 100 (highest quality). We expect the reference to be rated close to 100, and the anchor close to 0. The MUSHRA evaluation aims to indicate whether all models are perceptually equivalent, meaning they all have a similar quality. Additionally, we want to verify whether the playing techniques influence the modeling quality to understand if a given model renders a better representation of percussive than sustained sounds, for example. For such, we use two-way ANOVA on the MUSHRA ratings using *model* as the first factor with 5 levels (REF, AC, SM+, eaQHM, and DDM) and *playing technique* as the second factor with three levels (sustained, percussive, and modulated). If at least one model is perceptually different from



(a) RMS Log Spectral Measure (↓).



(b) Itakura-Saito Divergence (↓).

Figure 2: Comparison of spectral modeling accuracy (↓) of DDM, eaQHM, SM+, and the anchor (AC) for sustained, percussive, and modulated musical instrument sounds. The top compares the median RMS-LSM across frames (↓ means lower is better) and the bottom compares the median ISD across frames. Table 2 lists the instruments and the corresponding abbreviations. The highlighted sounds were used in the perceptual evaluation. Note that *ordinario* is not indicated and also that eaQHM failed to converge for **Gs**.

the others, we wish to know which one and whether it is of higher or lower quality. We perform a *post hoc* analysis via pairwise comparisons using the Holm-Bonferroni method to control for type 1 errors (false positives). Additionally, we want to determine if the models are perceptually transparent (i.e., distinguishable from the reference), so we will compare their ratings with those of the reference.

4. RESULTS AND DISCUSSION

This section presents the results of the objective evaluation, of the listening test, and then compares them.

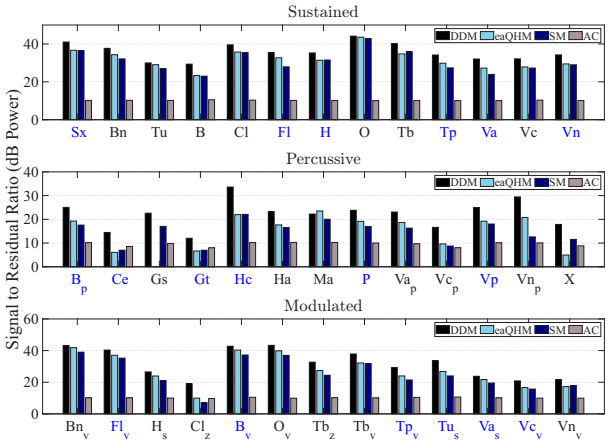


Figure 3: Comparison of temporal modeling accuracy (↑) of DDM, eaQHM, SM+, and the anchor (AC) for sustained, percussive, and modulated musical instrument sounds. The figure compares total SRR (↑ means higher is better). Table 2 lists the instruments and the corresponding abbreviations. The highlighted sounds were used in the perceptual evaluation. Note that *ordinario* is not indicated and also that eaQHM failed to converge for **Gs**.

4.1. Objective Evaluation Results

Fig. 2 presents the comparison of the spectral modeling measures RMS-LSM in Fig. 2a and ISD in Fig. 2b. Fig. 3 shows the comparison of the temporal modeling measure SRR. In both figures, the top row shows the results for *sustained*, the middle row for *percussive*, and the bottom row for *modulated* sounds (see Table 2 for more information). RMS-LSM and ISD are akin to distances, hence the symbol ↓ indicates that smaller values reflect better spectral modeling. On the other hand, the symbol ↑ associated with SRR indicates that higher values reflect better temporal modeling.

For sustained instrument sounds, Fig. 2a shows that DDM and eaQHM have comparable RMS-LSM in most cases, whereas SM+ is almost always higher. However, Fig. 2b suggests that DDM and SM+ have comparable ISD for most sustained sounds while eaQHM is generally lower. Finally, for percussive and modulated sounds, Fig. 2a and Fig. 2b show that both RMS-LSM and ISD for DDM and SM+ are higher than for eaQHM in most cases. Note that Fig. 3 shows that SRR for DDM is consistently higher for all sounds (except for **Ma**), whereas SM+ and eaQHM have comparable SRR in most cases, revealing that DDM captures more oscillatory energy with AM-FM sinusoids than either eaQHM or SM+.

The differences in RMS-LSM and ISD between DDM and SM+ are due both to peak selection (different peaks might be rejected, see Sec. 2.4) and to the estimation procedure, since SM+ models spectral peaks as QSS and DDM as nonstationary AM-FM sinusoids (see Table 1). However, the differences in RMS-LSM and ISD when contrasting DDM and SM+ with eaQHM can be attributed mainly to peak selection. For example, frames with spectral energy below the absolute threshold ρ_a across the entire frequency range will not be modeled by either DDM or SM+, resulting in high values for both RMS-LSM and ISD. For these frames, eaQHM will model the spectral peaks closest in frequency to the adaptive harmonic template as nonstationary sinusoids regardless of whether these correspond to noise instead. Note, however, that the SRR for eaQHM is still lower, indicating that modeling noise

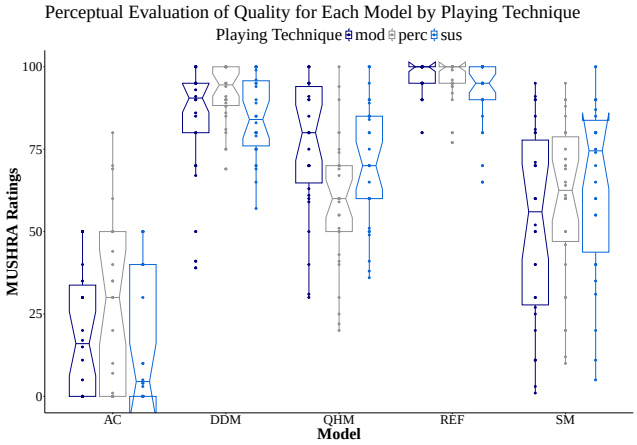


Figure 4: Comparison of perceptual evaluation of quality for the anchor (AC), DDM, eaQHM, the reference (REF), and SM+ for the playing techniques modulated (mod), percussive (perc), and sustained (sus).

below -110 dB does not improve modeling accuracy. Indeed, DDM has generally higher values of SRR than eaQHM and lower RMS-LSM and ISD, suggesting that, even if DDM models fewer spectral peaks than eaQHM, DDM models them more accurately leaving a residual more akin to instrumental noise.

4.2. Perceptual Evaluation Results

The MUSHRA test uses a *within-subjects* experimental design, where the within-subjects factors *model* and *playing technique* are completely crossed, resulting in a balanced design where each *model* and *playing technique* combination is evaluated 30 times. Thus, we used two-way repeated measures ANOVA to analyze the effect of *model* and *playing technique* on perceptual *quality*. The null hypothesis is that there is no difference in quality across models. However, we expect DDM and eaQHM to be better than SM+ and the anchor (AC) to be worse than all models, so we expect to find a statistically significant effect of *model* on *quality* ratings. Additionally, we do not expect the *playing technique* to impact the modeling quality, so we do not expect to find a statistically significant effect of *playing technique* on *quality*.

In total, 13 participants completed the test, and 8 participants who rated the hidden reference (REF) below 90 or the anchor (AC) above 90 for more than 15% of the references were rejected [53]. Thus, the analysis uses data from the 5 remaining participants. Fig. 4 shows the boxplot of the MUSHRA ratings for each model grouped by playing technique. The ratings were tested for extreme outliers, and only the rating for *Ce* (see Table 2) as the REF by one participant was removed for being considerably lower than the others. Additionally, the ratings were also tested for normality of residuals and sphericity to check that the two-way repeated measures ANOVA model is valid. The Shapiro-Wilk normality test found no evidence that the residuals are not normal ($W = 0.995$,$p = 0.092$).

The two-way repeated measures ANOVA showed a statistically significant main effect of *model* on *quality* ratings ($F(4, 16) = 40.98$, $p = 3.24 \times 10^{-8}$), but no significant main effect of *playing technique* ($F(2, 8) = 0.263$, $p = 0.775$) on *quality*. There was a statistically significant interaction effect between *model* and *playing technique* ($F(8, 32) = 10.94$, $p = 2.83 \times 10^{-7}$), suggesting that the main effect of *model* also depends on *playing technique*.

4.3. Post-Hoc Analysis

Further post-hoc tests were performed to determine if *playing technique* has an effect in *quality rating* for each isolated *model*. Additionally, multiple pairwise comparisons (corrected via Holm-Bonferroni) were performed to determine which interaction effects are significant.

4.3.1. Effect of Playing Technique

Firstly, we ran separate one-way repeated measures ANOVA analyses of the effect of *playing technique* on the *quality* ratings for each *model*. No statistically significant effects were found for the reference REF ($F(2, 8) = 2.714$, $p = 0.126$), the anchor AC ($F(2, 8) = 2.37$, $p = 0.156$), or DDM ($F(2, 8) = 3.643$, $p = 0.075$). However, somewhat surprisingly, statistically significant effects were found for eaQHM ($F(2, 8) = 25.86$, $p = 3.22 \times 10^{-4}$) and SM+ ($F(2, 8) = 20.31$, $p = 7.33 \times 10^{-4}$). Note that there is strong evidence of an effect of *playing technique* for eaQHM and SM+, revealing that the modeling quality from DDM is more consistent across playing techniques.

4.3.2. Pairwise Model Comparisons

Finally, we performed pairwise model comparisons using the t-test with Holm-Bonferroni corrected p-values. Two-sided comparisons yielded statistically significant differences between all models. So, we performed sequential right-tailed t-tests to determine which models have quality ratings significantly larger than others. Table 3 summarizes the results of the pairwise comparisons, where each row shows the result of the right-tailed comparison between the row model with the corresponding column model. Thus, the quality of REF is significantly greater than all other models, although evidence of the difference between REF and DDM is much smaller than the other models. Similarly, the quality of DDM is significantly greater than all other models, although the evidence of the difference between DDM and eaQHM is slightly smaller than the others. Additionally, the quality of eaQHM is significantly greater than that of the anchor (AC) and of SM+, albeit the evidence for the difference between eaQHM and SM+ is relatively much smaller. Finally, the quality of SM+ is significantly larger than the anchor (AC).

Table 3: Results of the post-hoc analysis. Pairwise comparisons between models after Holm-Bonferroni correction. The table should be read by rows to see the p-value for the right-tailed comparison between the model in the row and the model in the column.

	AC	**SM+**	**eaQHM**	**DDM**
REF >	$p < 2 \times 10^{-16}$	$p < 2 \times 10^{-16}$	$p < 2 \times 10^{-16}$	$p = 9.1 \times 10^{-3}$
DDM >	$p < 2 \times 10^{-16}$	$p < 2 \times 10^{-16}$	$p = 2.7 \times 10^{-9}$	
eaQHM >	$p < 2 \times 10^{-16}$	$p = 2.0 \times 10^{-3}$		
SM+ >	$p < 2 \times 10^{-16}$			

5. CONCLUSIONS

This work described the application of the distribution derivative method (DDM) in the estimation of the parameters of polynomial modulation sinusoids (PMS) used to model acoustic musical instrument sounds with nonstationary sinusoids. We have briefly reviewed the mathematical derivation of DDM and its use with the *windowed* discrete Fourier transform (DFT). The evaluation compared the performance of DDM with SM+ and eaQHM in modeling 39 sustained, modulated, and percussive sounds from different acoustic musical instruments and showed that DDM captures more oscillatory energy with AM-FM sinusoids. A MUSHRA listening test with 18 selected sounds confirmed that the modeling quality of DDM is higher than eaQHM and SM+. DDM is simple to implement and only requires the FFT and the pseudo-inverse to achieve state-of-the-art modeling performance.

Future work could investigate the application of DDM to model speech and polyphonic audio. Additionally, it would be beneficial to investigate the modeling performance of DDM with different windows because of how other windows concentrate spectral energy in the main lobe differently. Peak selection is particularly important for sinusoidal modeling with DDM because parameter estimation with eq. (13) using spectral peaks that do not correspond to sinusoids is currently the major source of modeling errors.

6. ACKNOWLEDGMENTS

The author would like to thank Qianyi (Rose) Sun for making the webpage for the online listening test.

7. REFERENCES

[1] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 34, no. 4, pp. 744–754, Aug 1986.

[2] X. Serra and J. O. Smith, "Spectral Modeling Synthesis: A sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Comp. Mus. J.*, vol. 14, pp. 12–24, 1990.

[3] K. Fitz and L. Haken, "Sinusoidal modeling and manipulation using lemur," *Comp. Mus. J.*, vol. 20, no. 4, pp. 44–59, 1996.

[4] Y. Ding and X. Qian, "Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (quasar) signal model," *J. Audio Eng. Soc.*, vol. 45, no. 7/8, pp. 571–584, 1997.

[5] A. S. Master and Y.-W. Liu, "Nonstationary sinusoidal modeling with efficient estimation of linear frequency chirp parameters," in *Proc. ICASSP*, vol. 5, 2003, pp. V–656.

[6] M. Caetano, G. P. Kafentzis, A. Mouchtaris, and Y. Stylianou, "Adaptive sinusoidal modeling of percussive musical instrument sounds," in *Proc. EUSIPCO*, 2013, pp. 1–5.

[7] W. Xue and M. Sandler, "Fast additive sinusoidal synthesis with a subband sinusoidal method," *IEEE Sig. Proc. Lett.*, vol. 20, no. 5, pp. 467–470, 2013.

[8] M. Caetano, G. P. Kafentzis, A. Mouchtaris, and Y. Stylianou, "Full-band quasi-harmonic analysis and synthesis of musical instrument sounds with adaptive sinusoids," *Appl. Sci.*, vol. 6, no. 5, p. 127, 2016.

[9] N. van Schijndel, M. Gomez, and R. Heusdens, "Towards a better balance in sinusoidal plus stochastic representation," in *Proc. WASPAA*, 2003, pp. 197–200.

[10] M. Caetano, G. P. Kafentzis, G. Degottex, A. Mouchtaris, and Y. Stylianou, "Evaluating how well filtered white noise models the residual from sinusoidal modeling of musical instrument sounds," in *Proc. WASPAA*, 2013, pp. 1–4.

[11] D. C. Rife and G. A. Vincent, "Use of the discrete Fourier transform in the measurement of frequencies and levels of tones," *Bell Sys Tech J*, vol. 49, no. 2, pp. 197–228, 1970.

[12] S. Kay, "A fast and accurate single frequency estimator," *IEEE Trans Acoust, Speech, Sig Proc*, vol. 37, no. 12, pp. 1987–1990, 1989.

[13] B. G. Quinn, "Estimating frequency by interpolation using Fourier coefficients," *IEEE Trans. Sig. Proc.*, vol. 42, no. 5, pp. 1264–1268, 1994.

[14] M. D. MacLeod, "Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones," *IEEE Trans. Sig. Proc.*, vol. 46, no. 1, pp. 141–148, 1998.

[15] Y. C. Xiao, P. Wei, X. C. Xiao, and H. M. Tai, "Fast and accurate single frequency estimator," *Electronics Letters*, vol. 40, no. 14, pp. 910–911, July 2004.

[16] E. Jacobsen and P. Kootsookos, "Fast, accurate frequency estimators [DSP tips & tricks]," *IEEE Sig. Proc. Mag.*, vol. 24, no. 3, pp. 123–125, May 2007.

[17] S. Provencher, "Estimation of complex single-tone parameters in the DFT domain," *IEEE Trans. Sig. Proc.*, vol. 58, no. 7, pp. 3879–3883, 2010.

[18] ——, "Parameters estimation of complex multitone signal in the DFT domain," *IEEE Trans. Sig. Proc.*, vol. 59, no. 7, pp. 3001–3012, July 2011.

[19] K. Duda, "DFT interpolation algorithm for Kaiser-Bessel and Dolph-Chebyshev windows," *IEEE Trans. Instr. Measure.*, vol. 60, no. 3, pp. 784–790, 2011.

[20] D. Belega and D. Petri, "Frequency estimation by two- or three-point interpolated Fourier algorithms based on cosine windows," *Sig. Proc.*, vol. 117, pp. 115–125, 2015.

[21] Ç. Candan, "Fine resolution frequency estimation from three DFT samples: Case of windowed data," *Sig. Proc.*, vol. 114, pp. 245–250, 2015.

[22] K. J. Werner and F. G. Germain, "Sinusoidal parameter estimation using quadratic interpolation around power-scaled magnitude spectrum peaks," *Appl. Sci.*, vol. 6, no. 10, p. 306, 2016.

[23] L. Fan and G. Qi, "Frequency estimator of sinusoid based on interpolation of three DFT spectral lines," *Sig. Proc.*, vol. 144, pp. 52–60, 2018.

[24] A. S. Master and K. Lee, "Explicit onset modeling of sinusoids using time reassignment," in *Proc. ICASSP*, vol. 3, 2005, pp. iii/221–iii/224 Vol. 3.

[25] J. J. Wells and D. T. Murphy, "High-accuracy frame-by-frame non-stationary sinusoidal modelling," in *Proc. DAFx*, 2006, pp. 253–258.

[26] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 19, no. 2, pp. 290–300, 2011.

[27] D. Fourer, F. Auger, and G. Peeters, "Local AM/FM parameters estimation: Application to sinusoidal modeling and blind audio source separation," *IEEE Sig. Proc. Lett.*, vol. 25, no. 10, pp. 1600–1604, 2018.

[28] M. Abe and J. Smith, "AM/FM rate estimation for time-varying sinusoidal modeling," in *Proc. ICASSP*, vol. 3, 2005, pp. iii/201–iii/204 Vol. 3.

[29] M. Betser, "Sinusoidal polynomial parameter estimation using the distribution derivative," *IEEE Trans. Sig. Proc.*, vol. 57, no. 12, pp. 4633–4645, 2009.

[30] M. Zivanovic and J. Schoukens, "On the polynomial approximation for time-variant harmonic signal modeling," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 19, no. 3, pp. 458–467, March 2010.

[31] B. Hamilton and P. Depalle, "A unified view of non-stationary sinusoidal parameter estimation methods using signal derivatives," in *Proc. ICASSP*, 2012, pp. 369–372.

[32] S. Muševič and J. Bonada, "Derivative analysis of complex polynomial amplitude, complex exponential with exponential damping," in *Proc. ICASSP*, 2013, pp. 488–492.

[33] O. Derrien, R. Badeau, and G. Richard, "Parametric audio coding with exponentially damped sinusoids," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 21, no. 7, pp. 1489–1501, 2013.

[34] J. Neri, P. Depalle, and R. Badeau, "Damped chirp mixture estimation via nonlinear bayesian regression," in *Proc. DAFx*, 2021, pp. 65–72.

[35] L. Girin, S. Marchand, J. Di Martino, A. Röbel, and G. Peeters, "Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals," in *Proc. WASPAA*, 2003, pp. 193–196.

[36] A. Röbel, "Frequency slope estimation and its application for non-stationary sinusoidal parameter estimation," in *Proc. DAFx*, 2007.

[37] B. Hamilton, P. Depalle, and S. Marchand, "Theoretical and practical comparisons of the reassignment method and the derivative method for the estimation of the frequency slope," in *Proc. WASPAA*, 2009, pp. 345–348.

[38] B. Hamilton and P. Depalle, "Comparisons of parameter estimation methods for an exponential polynomial sound signal model," in *Proc. AES Convention*, 2012, pp. 6–3.

[39] S. Marchand and P. Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling," in *Proc. DAFx*, 2008.

[40] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Sig. Proc.*, vol. 43, no. 5, pp. 1068–1089, 1995.

[41] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 360–371, 2006.

[42] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An extension of the adaptive quasi-harmonic model," in *Proc. ICASSP*, 2012, pp. 4605–4608.

[43] L. Schwartz, *Théorie des Distributions (In French)*. Publications de l'Institut de Mathématique de l'Université de Strasbourg, Actualités Scientifiques et Industrielles, no. 9, 1950, vol. I.

[44] ——, *Théorie des Distributions (In French)*. Publications de l'Institut de Mathématique de l'Université de Strasbourg, Actualités Scientifiques et Industrielles, no. 10, 1951, vol. II.

[45] G. Peeters and X. Rodet, "Signal characterization in terms of sinusoidal and non-sinusoidal components," in *Proc. DAFx*, Barcelonne, Spain, Nov. 1998, pp. 1–1.

[46] M. Lagrange, S. Marchand, and J.-B. Rault, "Sinusoidal parameter extraction and component selection in a non stationary model," in *Proc. DAFx*, 2002.

[47] M. Zivanovic, A. Röbel, and X. Rodet, "A new approach to spectral peak classification," in *Proc. EUSIPCO*, 2004, pp. 1277–1280.

[48] A. Röbel, M. Zivanovic, and X. Rodet, "Signal decomposition by means of classification of spectral peaks," in *Proc. ICMC*, 2004.

[49] M. Zivanovic, "Detection of non-stationary sinusoids by using joint frequency reassignment and null-to-null bandwidth," *Dig. Sig. Proc.*, vol. 21, no. 1, pp. 77–86, 2011.

[50] S. Kraft, A. Lerch, and U. Zölzer, "The tonalness spectrum: feature-based estimation of tonal components," in *Proc. DAFx*, 2013.

[51] M. Caetano and P. Depalle, "On the estimation of sinusoidal parameters via parabolic interpolation of scaled magnitude spectra," in *Proc. DAFx*, 2021, pp. 81–88.

[52] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 24, no. 5, pp. 380–391, 1976.

[53] R. BS-1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," ITU-R, Tech. Rep., 2015.

[54] M. Caetano, "Morphing musical instrument sounds with the sinusoidal model in the sound morphing toolbox," in *Perception, Representations, Image, Sound, Music*, R. Kronland-Martinet, S. Ystad, and M. Aramaki, Eds. Cham: Springer International Publishing, 2021, pp. 481–503.