APPENDIX

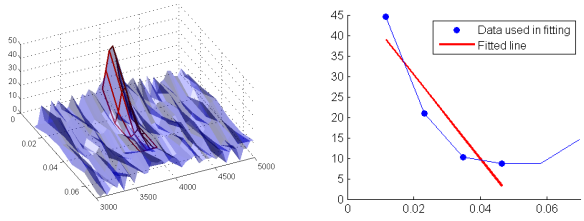
A. COMPARISON WITH PREVIOUS FEATURE EXTRACTION METHODS

A.1 Comparison with SMS

The Spectral Modeling Synthesis (SMS) method [Serra and Smith III 1990] detects a peak also in the power spectrogram, tracks the one peak point over time, and forms an amplitude envelope. One can certainly use this amplitude envelope to infer the damping value, for example, by linear regression of the logarithmic amplitude values (which is the approach adopted by [Välimäki et al. 1996]). There are, however, several disadvantages of this approach.

First of all, tracking only the peak point over time implies that the frequency estimation is only accurate to the width of the frequency bins of power spectrogram. For example, for a window size of 512 samples, the width of a frequency bin is about 86 Hz, direct frequency peak tracking has frequency resolution as coarse as 86 Hz.

Serra and Smith pointed out this problem [Serra and Smith III 1990], and proposes to improve the accuracy by taking the two neighboring frequency bins around the peak and performing a 3-point curve fitting to find the real peak [Serra 1989]. Our method takes a further step: instead of 3 points per time frame, we use all points within a rectangular region. The region extends as far as possible in both frequency and time axes until (a) the amplitude falls under a threshold to the peak amplitude, or (b) a local minimum in amplitude is reached. We then use an optimizer to find a damped sinusoid whose power spectrogram best matches the shape of the input data in the region of interest. An example is shown in Fig. 15a, where the blue surface is the power spectrogram of the input sound clip, and the overlay red mesh is the power spectrogram of the best fitted damped sinusoid.

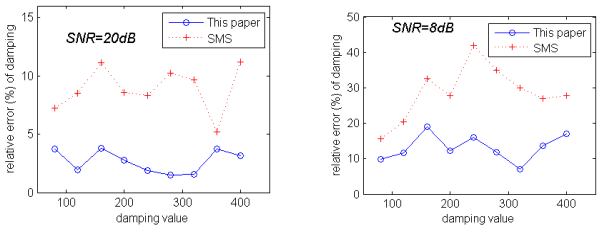


(a) (b)

Fig. 15: Estimation of damping value in the presence of noise, using (a) our local shape fitting method and (b) SMS with linear regression.

Secondly, for linear regression to work well, there must be at least two points (the more the better) along the time axis, before the signal falls to the level of background noise. For high damping values, there will be only a few data points along the time axis. On the other hand, we know that the damping value is also reflected in the width of the hill, so when there are not enough points along the time axis, there are more points along the frequency axis with significant heights–which will help determining the damping value in our surface fitting method.

Taking more points into account makes it less sensitive to noise. In Fig. 16, we simulated a noisy case where white noise with signal-to-noise ratio (SNR)=8 dB is added to a damped sinusoid with damping value 240, and use (a) our local surface fitting method and (b) SMS with linear regression to infer the damping value. In this particular example, due to the high damping value and high noise level, only 4 points participate in linear regression, while 24 points are considered in our method. Our shape fitting is less sensitive to irregularities than the fitted line in SMS. The average damping error versus damping value for both methods are plotted in Fig. 16a and Fig. 16b, where SNR=20 dB and 8 dB respectively.



(a) (b)

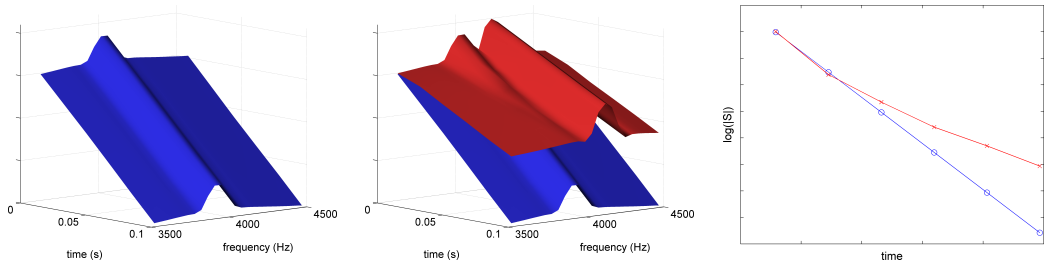
Fig. 16: Average damping error versus damping value for our method and SMS.

Mathematically, the 2D power spectrogram contains as much information as the original time domain signal (except for the windowing effect and the loss of phase). Using only a 1D sequence inevitably discards a portion of all available information (as in SMS), and in some cases (e.g. high damping values and high noise level) this portion is significant. Our surface matching method utilizes as much information as possible. Fitting a surface is indeed more costly than fitting a line, but it also achieves higher accuracy.

A.2 Comparison with a Phase Unwrapping Method

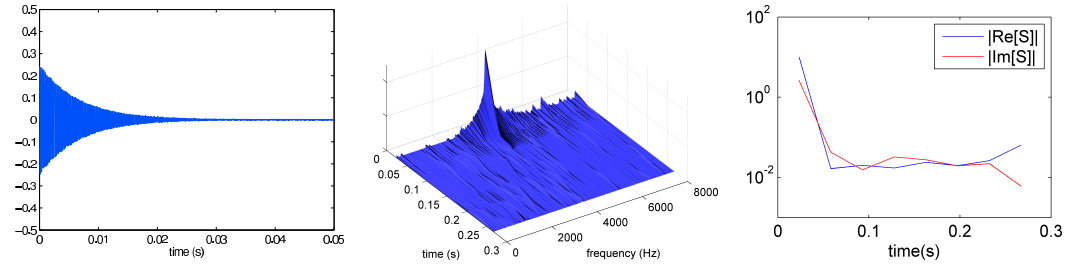
The 'phase unwrapping' technique proposed by [Pai et al. 2001] and [Corbett et al. 2007] is known for its ability to separate close modes within one frequency bin. Our method, however, works under a different assumption, and the ability to separate modes within a frequency bin has different impacts in our framework and theirs. In their framework, the extracted features $\{f_i, d_i, a_i\}$ are directly used in the sound synthesis stage and thus control the final audio quality. In our case, the features are only used to guide the subsequent parameter estimation process. In this process, two close modes will show up as near-duplicate points in the (f, d)-space. Because as pointed out by [Pai et al. 2001], modes with close frequencies usually result from the shape symmetry of the sounding object, and their damping values should also be close. In the process of fitting material parameters, or more specifically, in computing the feature domain metric, replacing these near-duplicate points with one point does not affect the quality of the result much.

Secondly, despite its ability to separate nearby modes, [Corbett et al. 2007] also proposes to merge modes if their difference in frequency is not greater than human's audible frequency discrimination limit (2-4 Hz). Among the multiple levels of power spectrograms that we used, the finest frequency resolution (about 3 Hz) is in fact around this limit.



(a) (b) (c)

Fig. 17: Interference from a neighboring mode located several bins away.



(a) (b) (c)

Fig. 18: A noisy, high damping experiment.

On the other hand, our proposed feature extraction algorithm offers some advantages and achieves higher accuracy compared with [Pai et al. 2001] and [Corbett et al. 2007] in some cases. When extracting the information of a mode, other modes within the same frequency bin (which are successfully resolved by the Steiglitz-McBride algorithm [Steiglitz and McBride 1965] underlying [Pai et al. 2001] and [Corbett et al. 2007]) are not the only source of interference. Other modes from several bins away also affect the values (complex or magnitude-only alike) in the current bin, known as the 'spillover effect'. In order to minimize this effect, the greedy method proposed in our paper collects the modes with the largest average power spectral density first. Therefore, when examining a mode, the neighboring modes that have higher energy than the current one are already collected, and their influence removed. This can be demonstrated in Fig. 17. The original power spectrogram of a mode (f_1, d_1, a_1) is shown in Fig. 17a. The values at the frequency bin F_k containing f_1 are plotted over time, shown as the blue curve in Fig. 17c. In Fig. 17b, the the presence of another strong mode (f_2, d_2, a_2) located 5 bins away changes the values at F_k, plotted as the red curve in Fig. 17c. The complex values of the STFT at F_k are not shown, but they are similarly interfered. If these complex values at F_k are directly fitted with the Steiglitz-McBride algorithm in the works by [Pai et al. 2001] and [Corbett et al. 2007], the estimated damping has a 20% error. The greedy approach in our multi-level algorithm removes the influence of the neighboring mode first, resulting in a 1% damping error.

Based on our experimentations, we also found that the universal frequency-time resolution used in [Pai et al. 2001] and [Corbett et al. 2007] is not always most suitable for all modes. Our method uses a dynamic selection of frequency-time resolution to address this problem. For example, in the case of high damping values, under a fixed frequency-time resolution, there may only be a few points above noise level along the time axis, which will undermine the accuracy of the Steiglitz-McBride algorithm. Fig. 18 shows such an example, the damping value ($150\ s^{-1}$) is high but not unreasonable, as shown in the time domain signal Fig. 18a, where a white noise with SNR=60 dB is added. The power spectrogram is shown in Fig. 18b. We implemented the method in the paper by [Corbett et al. 2007] using the suggested 46 ms window size (with $N_{overlap} = 4$) and tested on the above case. The input to this method is the complex values at the peak frequency bin, whose magnitudes of the real and imaginary parts are shown in Fig. 18c, and an error of 5.7% for damping is obtained. As a comparison, our algorithm automatically selects a 23 ms window size and fits the local shape in a 6×5 region in the frequency-time space, yielding merely a 0.9% error for damping.

B. CONSTANTS AND FUNCTIONS

We provide here the actual values and forms used in our implementation for the constants and functions introduced in Sec. 5.2,

For the relationship between critical-band rate z (in Bark) and frequency (in Hz), we use

$$Z(f) = 6\sinh^{-1}(f/600) \tag{30}$$

that approximates the empirically determined curve shown in Fig. 4a [Wang et al. 1992].

We use $c_z = 5.0$ and $c_d = 100.0$ in Eqn. 21 and Eqn. 22.

In Eqn. 23, the weight w_i associated to a reference feature point ϕ_i is designed to be related to the energy of mode i. The energy can

be found by integrating the power spectrogram of the damped sinusoid, and we made a modification such that the power spectrogram is transformed prior to integration. The image domain transformation introduced in Sec. 5.2.1, which better reflects the perceptual importance of a feature, is used.

The weight $\tilde{u}_{ij}$ used in Eqn. 24 is $\tilde{u}_{ij} = 0$ for $k(\phi_i, \tilde{\phi}_j) = 0$, and $\tilde{u}_{ij} = 1$ for $k(\phi_i, \tilde{\phi}_j) > 0$ (u_{ij} is defined similarly).

For the point-to-point match score $k(\phi_i, \tilde{\phi}_j)$ in Eqn. 24, we use

$$k(\phi_i, \tilde{\phi}_j) = k(D) = \begin{cases} 1.0 - 0.5D & \text{if } D \leq 1.0 \\ 0.5/D & \text{if } 1.0 < D \leq 5.0 \\ 0 & \text{if } 5.0 < D \end{cases} \quad (31)$$

where $D = D(\phi_i, \tilde{\phi}_j)$ is the Euclidean distance between the two feature points (Eqn. 20).

REFERENCES

CORBETT, R., VAN DEN DOEL, K., LLOYD, J. E., AND HEIDRICH, W. 2007. Timbrefields: 3d interactive sound models for real-time audio. *Presence: Teleoperators and Virtual Environments 16,* 6, 643–654.

PAI, D. K., DOEL, K. V. D., JAMES, D. L., LANG, J., LLOYD, J. E., RICHMOND, J. L., AND YAU, S. H. 2001. Scanning physical interaction behavior of 3d objects. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. SIGGRAPH '01. ACM, New York, NY, USA, 87–96.

SERRA, X. 1989. A system for sound Analysis/Transformation/Synthesis based on a deterministic plus stochastic decomposition. Ph.D. thesis.

SERRA, X. AND SMITH III, J. 1990. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal 14,* 4, 12–24.

STEIGLITZ, K. AND MCBRIDE, L. 1965. A technique for the identification of linear systems. *Automatic Control, IEEE Transactions on 10,* 4, 461–464.

VÄLIMÄKI, V., HUOPANIEMI, J., KARJALAINEN, M., AND JÁNOSY, Z. 1996. Physical modeling of plucked string instruments with application to real-time sound synthesis. *Journal of the Audio Engineering Society 44,* 5, 331–353.

WANG, S., SEKEY, A., AND GERSHO, A. 1992. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications 10,* 5 (June), 819–829.