A Comparison of Extended Source-filter Models for Musical Signal Reconstruction

Cheng, T; Dixon, S; Mauch, M

A COMPARISON OF EXTENDED SOURCE-FILTER MODELS FOR MUSICAL SIGNAL RECONSTRUCTION

Tian Cheng, Simon Dixon, Matthias Mauch†*

Centre for Digital Music
Queen Mary University of London
London, United Kingdom
{t.cheng, s.e.dixon, m.mauch}@qmul.ac.uk

ABSTRACT

Recently, we have witnessed an increasing use of the source-filter model in music analysis, which is achieved by integrating the source filter model into a non-negative matrix factorisation (NMF) framework or statistical models. The combination of the source-filter model and NMF framework reduces the number of free parameters needed and makes the model more flexible to extend. This paper compares four extended source-filter models: the source-filter-decay (SFD) model, the NMF with time-frequency activations (NMF-ARMA) model, the multi-excitation (ME) model and the source-filter model based on β-divergence (SFbeta model). The first two models represent the time-varying spectra by adding a loss filter and a time-varying filter, respectively. The latter two are extended by using multiple excitations and including a scale factor, respectively. The models are tested using sounds of 15 instruments from the RWC Music Database. Performance is evaluated based on the relative reconstruction error. The results show that the NMF-ARMA model outperforms other models, but uses the largest set of parameters.

1. INTRODUCTION

The source-filter model is a widely-used approximate physical model (considered as a physical model only when the coupling between the source and the filter is weak) for musical instrument modelling. The source (also called excitation) represents the vibrating object, and the filter models the frequency response of the instrument body. Introductions to physical modelling and source-filter models can be found in [1] and [2], respectively. Since Virtanen et al. used the source-filter model in audio analysis, and estimated the model parameters using the methods extended from NMF and non-negative matrix deconvolution (NMD) [3], we have observed some combinations of the source-filter model and NMF frameworks or statistical models for music analysis. In these combined models, the parameters are estimated in NMF framework [3, 4, 5, 6, 7, 8], with NMD [9], non-negative tensor factorisation [10], statistical models (Gaussian Scaled Mixture Model and extended Instantaneous Mixture Model) [11], or using EM [12].

By using the source-filter model, the spectral basis can be represented as a product of a source and filter, which reduces the number of free parameters and makes the estimation more reliable. On the other hand, the NMF or statistical model not only provides the baseline update rules for estimating parameters, but also makes the model more flexible to extend. For instance, the multi-excitation model extended the excitation as a weighted sum of multiple excitations with a harmonic constraint [8]; and Kirchhoff et al. extended the source-filter model with a scaling factor to compensate for gain differences [9]. A further extension is to include the time dimension in order to describe the time-varying spectral energy distribution. The source-filter-decay model proposed by Klapuri [13] extended the source-filter model with a loss filter to represent the time-varying spectral shape of pitched musical instruments. In [14], a model was proposed for representing the time-varying spectral characteristics of a quasi-harmonic instrument sound by assuming the spectral envelope to be determined by the partials' amplitude trajectories. A source-filter factorisation was proposed to model non-stationary audio events in [15]. In this model, the source works as the spectral basis of the NMF, and the filter is extended and works as the frequency-dependent temporal activations. The parameters of the source-filter model are estimated based on NMF. These models have shown their usefulness in several MIR applications, such as source separation [6, 10, 12, 13], melody extraction [5, 11] and music transcription [4, 7, 8, 9].

In this paper, four extended models are chosen for comparison: the source-filter-decay (SFD) model [13], the NMF with time-frequency activations (NMF-ARMA) model [15], the multi-excitation (ME) model [8] and the source-filter model based on β-divergence (SFbeta model) [9]. For the sake of completeness, a standard NMF is also included as a base line of comparison. The evaluation is based on the relative reconstruction error, and the complexities of the models are analysed in terms of parameter dimensionality. The results tested on the sounds from 15 instruments show that the average relative reconstruction error of the NMF-ARMA model is smallest, while its parameter dimensionality is largest. It approximates wind instruments well, while the other three models have better results on string instruments. All the models perform well on piano and guitar, while no model works well for all the instruments because of differences between the structures of the instruments. The poor performance for violin and vocals indicates a limitation of the models when encountering vibrato, which can be investigated in future work.

The rest of the paper is organised as follows: Section 2 gives a brief introduction to the models with modifications and parameters. The comparison results are illustrated in Section 3. Conclusions are drawn in Section 4.

* Tian Cheng is supported by a China Scholarship Council (CSC)/ Queen Mary Joint PhD scholarship.

† Matthias Mauch is funded by a Royal Academy of Engineering Research Fellowship.

2. MODELS

In this section, we present four different extended source-filter models. We mainly focus on the motivations and how the models are formulated rather than the detailed parameter learning equations which can be found in corresponding papers. Where applicable, we specify the modifications we made in order to run the methods and provide information on parameter settings.

2.1. Source-Filter-Decay Model

The source-filter-decay model [13] provides a way of representing the time-varying spectral energy distribution of pitched musical instruments. The changing spectral distribution of an instrument is modelled by extending the source-filter model by a loss filter, which models the frequency-dependent decay along the time axis. The model on a decibel scale is given as follows:

$$S^{(t)}_{\mathrm{dB}}(f_h) = \gamma_{\mathrm{dB}} + X_{\mathrm{dB}}(h) + B_{\mathrm{dB}}(f_h) + tL_{\mathrm{dB}}(f_h) + E^{(t)}_{\mathrm{dB}}(f_h) \tag{1}$$

where $f_h \approx hF$, is the frequency of h^{th} harmonic of the fundamental frequency F, $S^{(t)}_{\mathrm{dB}}(f_h)$ is the power spectrum (but only modelled at the positions of the harmonics), γ_{dB} denotes the overall gain of the sound, $X_{\mathrm{dB}}(h)$ is the initial level of the h^{th} harmonic, $B_{\mathrm{dB}}(f_h)$ represents the frequency response of the instrumental body, $L_{\mathrm{dB}}(f_h)$ is the frequency-dependent loss filter and $E^{(t)}_{\mathrm{dB}}(f_h)$ represents modelling error.

The 'source' X, 'filter' B and 'decay' L are further represented by the linear models:

$$\begin{aligned} X_{\mathrm{dB}}(h) &= \sum_{i=1}^{C_x} \xi_i x_i(h) \\ B_{\mathrm{dB}}(f) = \sum_{j=1}^{C_b} \beta_j b_j(f) &\text{ and } L_{\mathrm{dB}}(f) = \sum_{k=1}^{C_l} \lambda_k l_k(f) \end{aligned} \tag{2}$$

The basis functions $x_i(h)$ are found by performing PCA on the harmonics of sounds collected from 33 instruments, while $b_j(f)$ and $l_k(f)$ are defined in the same way with overlapped triangular bandpass responses on a critical-band frequency scale. After choosing the basis functions, $X_{\mathrm{dB}}(h)$, $B_{\mathrm{dB}}(f_h)$ and $L_{\mathrm{dB}}(f_h)$ are determined by the weights ξ_i, β_j and λ_k, respectively.

The parameters are estimated by minimizing the least-square (LS) error between the observed and modelled harmonic level using a weighted LS estimator. The influence of γ_{dB} is eliminated by performing subtraction between two observed harmonics or two consecutive frames.

2.1.1. Modifications

As the F0 estimation method used in the model [16] is unavailable to us, we use the pitches extracted by the SWIPE algorithm [17] with manual corrections (referred to as detected F0s, also used in Section 2.3.1 and 2.4.1). The model is built on the first two frames with stable pitches (frames after transient) of each note.

As the model only captures the harmonic levels of the sound, we convolve the result with the magnitude response of the window function to generate the reconstruction.

2.1.2. Parameters

The model is analysed in two scenarios: with and without decay filter, denoted by *SFD*(111) and *SFD*(110), respectively. However, as the decay rate modelled in two frames is not reliable to reconstruct the spectra of the whole note clip, we use only the model without decay filter for the reconstruction.

2.2. NMF-ARMA Model

Hennequin et al. extended the temporal activations of the standard NMF framework to be frequency-dependent, in order to model non-stationary notes [15]. The spectral basis and frequency-dependent activations in the NMF framework work as the sources and time-varying filters in the source-filter model. The time-varying filters are modelled using the Autoregressive Moving Average (ARMA) model and parameters are learned in the NMF framework, which is called source-filter factorisation. The spectrogram is modelled as follows:

$$V_{ft} \approx \hat{V}_{ft} = \sum_{r=1}^{R} \omega_{fr} h_{rt}(f) \tag{3}$$

where V_{ft} and $\hat{V}_{ft}$ are the original and reconstructed spectrograms, ω_{fr} are the spectral bases (the sources), $h_{rt}(f)$ are the frequency-dependent activations (the time-varying filters), which are parameterized following the general ARMA model:

$$h_{rt}(f) = \delta^2_{rt} \frac{|\sum_{q=0}^{Q} b^q_{rt} e^{-i2\pi v_f q}|^2}{|\sum_{p=0}^{P} a^p_{rt} e^{-i2\pi v_f p}|^2} \tag{4}$$

where δ^2_{rt} is the global gain of the filter, and b^q_{rt} and a^p_{rt} are the coefficients of the MA and AR parts of the filter, respectively. $v_f = (f-1)/(2(F-1))$, where f is frequency bin and F the total number of frequency bins.

This time-varying filter represents the spectral variations of the sound which are not modelled in standard NMF. The parameters are learned in an NMF framework using β-divergence.

2.2.1. Parameters

For each instrument, N sources are used, one for each note. Two sets of ARMA parameters are in use: $Q = 0, P = 2$ for the instruments with strongly varying spectral shapes and $Q = 1, P = 1$ for others. They are represented by *ARMA*(02) and *ARMA*(11), respectively.

2.3. Multi-Excitation Model

The multi-excitation model is motivated by the non-smooth structure of the spectral envelopes often observed in wind instruments [8]. To tackle this problem, note-varying excitations are represented by the weighted summation of excitation bases, which are modelled under a harmonic constraint as follows:

$$\begin{aligned} e_{n,j}(f) &= \sum_{m=1}^{M} a_{m,n,j} G(f - m f_0(n)) \\ a_{m,n,j} &= \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} \end{aligned} \tag{5}$$

where $e_{n,j}(f)$ is the excitation for pitch n and instrument j, $a_{m,n,j}$ is the amplitude of the m^{th} partial of the same note and $G(f - mf_0(n))$ is the harmonic component of pitch n. $v_{i,m,j}$ is the excitation basis vector belonging to partial m and instrument j, and $w_{i,n,j}$ is the weight of the i^{th} excitation basis for pitch n and instrument j.

The spectral basis function is modelled as the product of the excitation and the filter h in the usual way:

$$b_{n,j} = h_j(f)e_{n,j}(f) \tag{6}$$

and the reconstructed spectrogram is given as follows:

$$\hat{x}_t(f) = \sum_{n,j} g_{n,t,j} b_{n,j} \tag{7}$$

where $g_{n,t,j}$ are gains of instrument j.

The parameters are learnt in an NMF framework with KL divergence. For post-processing, temporal continuity is enforced over the gains by adding a cost term to penalize large changes in the gains between adjacent frames.

2.3.1. Modifications

The harmonic components are built based on detected F0s rather than the ideal pitches. We give up temporal continuity as no significant improvement is found (maybe because of an unsuitable parameter). Instead, we apply the sparsity constraint used in [18] for the post-processing as the test dataset only consists of isolated notes.

2.3.2. Parameters

The system is tested with 1,2 and 4 excitations (represented by *ME(I)*, where I is the number of excitations) to find out the relation between the number of excitations and the performance. In this paper, only the situation with one instrument at a time has been considered.

2.4. SFbeta Model

The source-filter model proposed by Kirchhoff et al. [9] is for estimating the missing templates for user-assisted music transcription. The model is built using a common excitation spectrum and a filter response on a log-frequency scale with a scaling factor. The proposed source-filter model represents the spectrum $\mathbf{w}_p$ of pitch ϕ_p as follows:

$$\mathbf{w}_p \approx \hat{\mathbf{w}}_p = \mathbf{s}_p \cdot \overset{\phi_p\downarrow}{\mathbf{e}} \otimes \mathbf{h} \tag{8}$$

where $\hat{\mathbf{w}}_p$ is the estimated spectrum, $\mathbf{s}_p$ is the scaling factor, $\mathbf{e}$ is the excitation, and $\mathbf{h}$ the filter response. The frequency is represented on a logarithmic scale. The $\otimes$ operator denotes element-wise multiplication of the vector, and the operator $\phi_p \downarrow$ shifts the excitation spectrum $\mathbf{e}$ along the frequency axis by ϕ_p frequency bins.

For all pitches ϕ_p ($p \in [1, \cdots, P]$), the scalars s_p are combined into a vector $\mathbf{s}$ of length P, and vectors $\mathbf{w}_p$ are combined into a matrix $\mathbf{W} \in R_+^{K,P}$, where K is the number of frequency bins. $\hat{\mathbf{W}}$ is a matrix with the same dimension as $\mathbf{W}$ combined from vectors of $\hat{\mathbf{w}}_p$.

The parameters $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{h}$ are estimated by using gradient descent on each vector iteratively to gradually decrease the β-divergence between $\mathbf{W}$ and $\hat{\mathbf{W}}$. The vectors are randomly initialized and details of the derivation of the update equations can be found in [19].

Table 1: *Instrument categories*

Categories	Instrument
String	piano, harpsichord, guitar, violin
Wind	accordion, harmonica, pipe organ, horn, saxophone, oboe, bassoon, clarinet, flute
Vocal	alto (female), tenor (male)

2.4.1. Modifications

In the model [9], spectra are shifted down to get the relative spectra according to the note pitches. Here we shift the spectra down using detected F0s rather than the ideal pitches, as not all instruments in the dataset are tuned to the same reference frequency. Preliminary tests have shown that this is necessary in order to obtain reasonable results.

3. EVALUATION

To evaluate a physical model for music analysis, the criterion is mainly based on the difference between the model's output and the original sound. In this paper, we evaluate the models according to the relative reconstruction error between the modelled and observed spectra. In addition, the parameter dimension of each model is analysed.

3.1. Evaluation Metrics

The relative reconstruction error (*RRE*) is chosen for the evaluation, which is defined as below:

$$RRE = \|OS - RS\|_F / \|OS\|_F \tag{9}$$

where $\|\cdot\|_F$ is the Frobenius norm, *OS* is the observed spectrum and *RS* is the reconstructed spectrum, both are amplitude spectra. We use the relative reconstruction error instead of the reconstruction error as the time-frequency representations of the models are different and the lengths of the sounds vary with instruments.

The parameter dimensionality indicates the complexity of the model, which is analysed in association with the time-frequency representation, the note ranges of the instruments, harmonic number and so on.

3.2. Experimental Setup

3.2.1. Test Dataset

To evaluate the four models, we choose the sounds of 15 instruments from the Musical Instrument Sound Database in the RWC Music Database[20], including string, wind instruments, female and male vocals, as listed in Table 1. Two violin recordings are chosen: Violin and Violin2 referring to notes played with and without vibrato, respectively.

For each instrument, we use the first 1s of each recorded note or the duration of the note if the note lasts for less than 1s. The onsets are detected using SuperFlux [21] with manual corrections. The F0s of the notes of the instruments are extracted using the SWIPE algorithm [17] with manual corrections. The ground truth (onsets and pitches) for these files can be found on-line.[1]

[1] available at https://code.soundsoftware.ac.uk/projects/onsetpitch/files

3.2.2. Time-Frequency Representation

For the source-filter-decay model, the NMF-ARMA model and the multi-excitation model, the original spectra are computed using the Short-Time Fourier Transform (STFT). Frames are segmented by a 2048-sample Hamming window with a hop-size of 441. A Discrete Fourier Transform is performed on each frame with 2-fold zero-padding. The sampling rate is $f_s = 44100$ Hz. For the SFbeta model, the time-frequency representation is calculated using a constant-Q transform [22] with 48 frequency bins per octave. The frequency range of all models is from 25 to 12500 Hz covering about 9 octaves.

The reconstructed spectra of the time-varying models (SFD and NMF-ARMA) cover the whole duration of the sound clips. For the multi-excitation model and SFbeta model, we first generate the spectral dictionary for the instruments based on the model, then calculate the reconstructed spectra using a standard NMF framework (multiplicative update) with the dictionary. This is also done when not using the decay filter in the source-filter-decay model.

3.3. Results

For the sake of completeness, the reconstruction result for NMF with no constraint is also included as a bottom line for comparison. The models are analysed in terms of the relative reconstruction error and parameter dimensionality.

3.3.1. Relative Reconstruction Errors

The results of the relative reconstruction error of the models are listed in Table 2. Models are tested with different parameters. Detailed parameters can be found in Section 2.

The average *RRE* of the source-filter-decay model is largest among the models, up to 42.9%. The model works relatively well on guitar, bassoon and flute (*RRE* $< 30\%$) but performs badly on the harmonica and vocals. Although the model is simplified by using a small set of parameters based on data from only 2 frames, the performance of the model is then affected.

The NMF-ARMA model outperforms other models with an average *RRE* of 16.9%. The model was proposed to model sounds with a strongly varying spectral shape, and the results show that it works well on most instruments (except violin with vibrato and vocals). The missing results (denoted by '–') are caused by inverting a singular matrix, which shows that the spectra of these instruments are flat and are not suitable for a model designed to deal with strong spectral variations. On the other hand, the improvement brought by using the parameter set $(0, 2)$ indicates that the notes have time-varying spectral shapes. An advantage of this model is manifested in the performance on the wind instruments with an average *RRE* of only 12.1%. Best results are found in bassoon and horn with *RRE*s of 4.99% and 6.10%, respectively. However, performance dramatically drops on violin with vibrato and vocals with about 40% *RRE*s using the parameter set $(1, 1)$, while the errors decrease by 4% for violin and by about 0.6% for the vocals using the parameter set $(0, 2)$. The poor performance occurs on the vibrato sounds such as those shown in Figure 1 (b). This is mainly because the model uses one filter per note, which fails to model the fluctuating pitches.

The multi-excitation model is proposed to approximate the non-smooth spectral envelopes of wind instruments by using a combination of excitations. We observe that the performance of

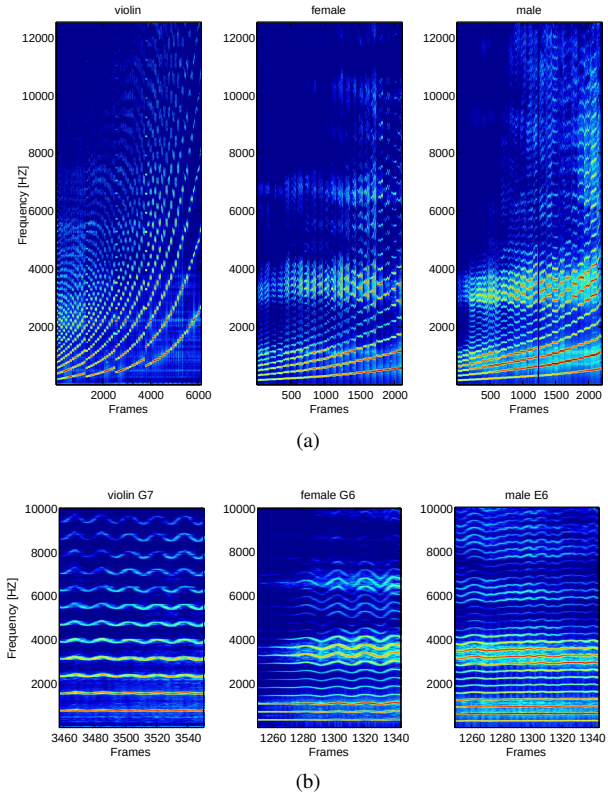

Figure 1: *Spectrograms of violin, female and male vocals: (a) all notes (64 notes for violin, 22 and 23 notes for female and male vocals, respectively), (b) individual note example.*

the model gradually improves with increasing number of excitations. The average *RRE* drops from 31.3% to 28.9% when using 2 excitations rather than 1; while a further 1.5% decrease is achieved by using 4 excitations. Notably, the errors fall by about 7% when employing 2 excitations on violin, bassoon and clarinet. The improvement by using more excitations indicates that the instrument has a non-smooth spectral envelope. The model works well on piano, pipe organ, guitar and flute even with one excitation. However, we also observe some noisy results when using more excitations in harpsichord, pipe organ, saxophone, flute and male voice.

The SFbeta model is the only model using a log-frequency scale among these models. The average *RRE* is 34.1% and the model is less sensitive to the choice of instrument. The model performs best on piano, guitar and tenor saxophone with *RRE*s of about 26%, while the worst results appear on clarinet, vocals, violin and bassoon. A notable phenomenon is that there are some inconsistencies of this model appearing on tenor saxophone, bassoon and flute, as we find that the other three models provide relatively poor results on tenor saxophone and perform well on flute. In addition, the best results of the source-filter-decay model and the NMF-ARMA model appear on bassoon, while the result on bassoon of the SFbeta model is one of the worst. By comparing the performance of the model on wind instruments, we find the model performs better on instruments with short and low frequency ranges.

Three out of four models, except the NMF-ARMA model, per-

Table 2: *Relative reconstruction errors (*RRE*), expressed as percentages. Results of piano and results better than that of piano are shown in bold. The symbol '–' means the result is not available, see text for details. The average error of the NMF-ARMA model is calculated using the better results of the two parameter sets.*

Instrument	MIDI range	*SFD*(110)	*ARMA*(11)	*ARMA*(02)	*ME*(1)	*ME*(2)	*ME*(4)	*SFbeta*	*NMF*
01 Piano	21-108	**31.7**	**7.43**	–	**11.4**	**10.7**	**10.7**	**26.6**	**5.34**
03 Harpsichord	28-88	52.1	16.4	–	19.7	19.0	22.4	31.2	7.69
06 Pipe Organ	36-91	42.9	13.1	–	15.1	15.5	15.4	36.8	9.19
07 Accordion	53-93	42.6	15.3	14.6	34.0	32.6	30.8	31.5	20.1
08 Harmonica	65-100	76.8	16.8	16.8	49.8	49.3	48.9	34.6	27.7
09 Guitar	40-76	**26.8**	11.7	–	12.7	**10.1**	**8.05**	**25.6**	**4.24**
15 Violin	55-101	37.2	40.4	36.4	38.0	31.1	29.9	38.5	15.4
15 Violin2	55-101	36.4	16.8	9.61	37.2	30.8	24.2	33.2	**4.92**
24 Horn	41-77	**31.5**	**6.10**	–	35.0	33.6	30.1	30.1	8.77
27 Tenor Sax	44-75	44.3	17.2	17.2	42.4	42.0	35.1	**25.7**	16.7
29 Oboe	58-91	44.6	8.23	–	22.8	20.2	19.1	35.2	14.3
30 Bassoon	34-72	**23.8**	**4.99**	–	32.3	25.8	23.7	38.5	9.13
31 Clarinet	50-89	48.5	15.6	–	51.7	44.9	42.2	43.7	20.9
33 Flute	60-96	**28.7**	12.2	–	14.6	14.6	15.3	37.2	9.15
46 Female	53-74	53.4	42.6	41.9	38.2	36.8	36.7	40.5	19.2
47 Male	53-74	64.8	38.4	37.9	45.8	45.2	45.7	37.2	22.1
Average		42.9	16.9		31.3	28.9	27.4	34.1	13.4
String Average		**36.8**	16.3		**23.8**	**20.3**	**19.1**	**31.0**	**7.52**
Wind Average		42.6	**12.1**		33.1	30.9	28.9	34.8	15.1
Vocal Average		59.1	39.9		42.0	41.0	41.2	38.9	20.7

form better on string instruments than on wind instruments, as shown in the average *RRE*s of string and wind instruments. We find that all models work well on piano and guitar, as a convincing evidence of the fitness of the source-filter model for these two instruments. On the other hand, all models perform badly on violin with vibrato and vocals. Two recordings of violin with vibrato and without vibrato are compared to find out whether the poor performance is caused by the vibrato. The *RRE*s of violin without vibrato (Violin2) are better than that of violin with vibrato (Violin) for all models. However, by checking the results of Violin2, a drop of 7.19% on *RRE* by using the parameter set $(0, 2)$ in the NMF-ARMA model shows that the violin sounds have a strong changing spectra distribution, while the improvement by using more excitations in the multi-excitation model indicates the non-smooth structure of the spectral envelope. So we could say that apart from the vibrato, the poor performance on violin is also caused by a changing spectral shape (as a result of, for instance, consistent changing pressure on the bow) and a non-smooth spectral envelope (4 strings). The bad results on vocals were not expected before the experiments, since the vocals have obvious filter responses as shown in Figure 1(a). And we found that the models capture the frequency response of the filter quite well in Figure 2. The frequency response corresponds to the vocal tract shape of the vowel /a:/ [23]. As no significant improvement is brought by using the parameter set $(0, 2)$ in the NMF-ARMA model and using more excitations in the ME model, the error is likely to stem from the reconstruction of the vibrato. The SFbeta model is least sensitive to vibrato. That is partly because for the constant-Q transform the frequency variations keep the same for all the partials, while the frequency variance gets larger at higher frequencies on the linear frequency scale. The SFbeta model performs worst on violin, bassoon and clarinet. They are exactly the same instruments which the multi-excitation model gets greatest improvement by using more excitations. This indicates the non-smooth spectral envelopes of these instruments and the utility of the multi-excitation model.

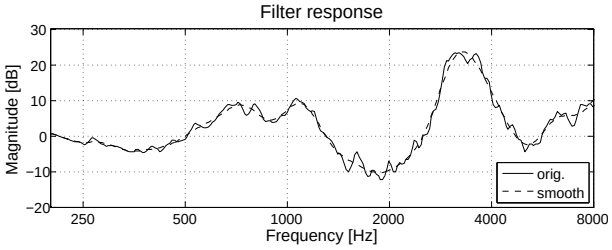

Figure 2: *Filter response of male vocal generated by the SFbeta model.*

3.3.2. Parameter Dimensionality

The parameter dimensions of the models are shown in Table 3. F and T are the numbers of frequency bins and time frames, respectively. The note range of each instrument is denoted by N. H is the number of the harmonics included in the model. I in the multi-excitation model indicates the number of excitations. To make it more intuitive, we list the dimensions for two instruments in Table 4, piano with 88 notes and harmonica with 20 notes.

The source-filter-decay model only has values at harmonic positions. The harmonic levels are represented by a weighted sum of C_H basis functions. The filter and decay are generated using a combination of C_B overlapped triangular bandpass filters. In this experiment, we use 15 basis functions and 20 bandpass filters, so only 55 parameters are used for each instrument in this model. When without the decay filter, the gains with NT parameters are also needed for the reconstruction.

The NMF-ARMA model builds each note using a source and

Table 4: *Parameter dimensions for piano and harmonica.*

Instrument	No.	SFD(111)	SFD(110)	ARMA	ME	SFbeta	NMF
Piano	88	55	7.4×10^5	2.3×10^6	7.4×10^5	5.9×10^6	9.2×10^5
Harmonica	20	55	3.8×10^4	1.4×10^5	4.0×10^4	3.1×10^5	7.9×10^4

Table 3: *Parameter dimensions.*

Algorithm	Dimension
SFD (111)	$C_H + 2C_B$
SFD (110)	$C_H + C_B + NT$
ARMA	$NF + NT(Q + P + 1)$
ME	$I(N + H) + F + NT$
SFbeta	$2F + NT$
NMF	$NF + NT$

a time-varying filter. The filter is represented by an ARMA model with $Q + P + 1$ parameters. So the number of parameters is $F + T(Q+P+1)$ per note. The parameter dimension increases linearly according to the note ranges of the instruments.

Both the multi-excitation model and the SFbeta model represent the spectra by multiplication of the dictionaries built by the models and the gains (NT). The source (excitation) of the multi-excitation model is a weighted sum of I excitation bases, and each basis is represented by H harmonics. The weights of each note is different, with NI weights in total. The filter is represented by F frequency bins. The whole model is represented by $I(N + H) + F + NT$ parameters.

For SFbeta model, the dictionary is generated by a source (F parameters) and filter (F parameters). The dimension of this model's parameters is $2F + NT$. The reason for the high figure of the model as shown in Table 4 is because the constant-Q transform has different numbers of frequency bins (F) and time frames (T). Apart from the influence of the TF representation, the parameter dimension of the SFbeta model is about the same as that of the multi-excitation model.

3.3.3. Comparison with NMF

With a large set of parameters, the average *RRE* of the NMF is smaller than that of all source-filter based models. Models with larger sets of parameters tend to have better results on the *RRE*. The NMF-ARMA model (with the largest set of parameters) outperforms the NMF on the average *RRE* of wind instruments. Besides reducing the number of free parameters, the source-filter models are employed because appropriate training data are not always available in real-world MIR applications and, as a result, pre-trained templates may not work [8].

4. CONCLUSIONS

In this paper, four extended source-filter models are evaluated according to the relative reconstruction error on sound clips from 15 instruments in the RWC Music Dataset. The results show that the source-filter-decay model captures the harmonic levels only with a small set of parameters, resulting in a large relative reconstruction error. The NMF-ARMA model obtains the smallest reconstruction result with the largest set of parameters. Performance is improved by using more excitations in the multi-excitation model, especially for violin, bassoon and clarinet, and the improvement indicates a non-smooth spectral envelope of the instrument. The results of the SFbeta model show low sensitivity to the choice of instrument. Overall, all the models perform well on piano and guitar, while no model works well for all the instruments because of differences between the structures of the instruments. The poor performance on vibrato indicates that a more flexible and shiftable structure is needed.

In future, we would like to develop a shiftable source-filter model for vibrato sounds using a constant-Q transform.

5. ACKNOWLEDGEMENTS

We thank Anssi Klapuri and Roland Badeau for generously sharing their code. We thank Holger Kirchhoff for making his code open-source.

6. REFERENCES

[1] V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen, "Discrete-time modelling of musical instruments," *Reports on Progress in Physics*, vol. 69, no. 1, pp. 1–78, Jan. 2006.

[2] D. Arfib, F. Keiler, U. Zölzer, and V. Verfaille, "Source-Filter Processing," in *DAFX: Digital Audio Effects: Second Edition*, pp. 279–320. J. Wiley Sons, Chichester, 2011.

[3] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, Hawaii, USA, 2006.

[4] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 109–112.

[5] J. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 169–172.

[6] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 327–332.

[7] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[8] J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. Canadas-Quesada, "Musical instrument sound

multi-excitation model for non-negative spectrogram factorization," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1144–1158, 2011.

[9] H. Kirchhoff, S. Dixon, and A. Klapuri, "Missing template estimation for user-assisted music transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 26–30.

[10] D. Fitzgerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[11] J. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[12] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and EM algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, pp. 5510–5513.

[13] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, 2007, pp. 53–56.

[14] H. Hahn, A. Röbel, J. Burred, and S. Weinzierl, "Source-filter model for quasi-harmonic instruments," in *International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 1–6.

[15] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[16] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *7th International Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, 2006, pp. 216–221.

[17] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speech and music.," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[18] E. Benetos and S. Dixon, "A shift-invariant latent variable model for automatic music transcription," *Computer Music Journal*, vol. 36, no. 4, pp. 81–94, 2012.

[19] H. Kirchhoff, S. Dixon, and A. Klapuri, "Derivation of update equations for a source-filter model based on beta-divergence," Tech. Rep., Queen Mary University of London, London, 2012.

[20] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music genre database and musical instrument sound database," in *4th International Society for Music Information Retrieval Conference (ISMIR)*, Baltimore, Maryland, USA, 2003, pp. 229–230.

[21] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *International Conference on Digital Audio Effects (DAFx)*, Maynooth, Ireland, 2013, pp. 1–7.

[22] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain, 2010.

[23] G. Bloothooft and R. Plomp, "Spectral analysis of sung vowels. III. Characteristics of singers and modes of singing," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 852–864, Mar. 1986.