

ANALYSIS OF MUSICAL INSTRUMENT SOUNDS BY SOURCE–FILTER–DECAY MODEL

Anssi Klapuri

Tampere University of Technology, Institute of Signal Processing
Korkeakoulukatu 1, FI-33720 Tampere, Finland

anssi.klapuri@tut.fi

ABSTRACT

This paper proposes a way of modelling the time-varying spectral energy distribution of musical instrument sounds. The model consists of an excitation signal, a body response filter, and a loss filter which implements a frequency-dependent decay. The three parts are further represented with a linear model which allows controlling the number of parameters involved. A method is proposed for estimating all the model parameters jointly, taking into account additive noise. The method is evaluated by measuring its accuracy in representing 33 musical instruments and by testing its usefulness in extracting the melodic line of one instrument from a polyphonic audio signal.

Index Terms— Music, modeling, least square methods, Viterbi decoding.

1. INTRODUCTION

This paper proposes a way of representing the timbre of pitched musical instruments. More specifically, the aim is to construct a model for the time-varying spectral energy distribution of sounds produced by a certain instrument. There are many uses for such models, including the recognition, coding, and synthesis of musical sounds.

By far the most widely used approach to modeling spectral energy distributions are Mel-frequency cepstral coefficients (MFCCs). Although these have several desirable properties due to their simplicity and perceptual and statistical properties, here we propose a more structured approach which leads to a better modeling accuracy for the diverse sound production mechanisms encountered in musical instruments. MFCCs and other “direct” ways of encoding the spectral shape have two flaws that are addressed here: it is hard to combine a good frequency resolution with pitch invariance, and secondly, some aspects of sound spectra are better described as a function of harmonic index instead of frequency, for example odd harmonics being stronger in certain wind instruments [1].

As a starting point here we adopt the source-filter model of sound production, where “source” represents a vibrating object such as a guitar string, and “filter” refers to the resonance structure of the rest of the instrument, which colors the produced sound. This framework has been used for decades in speech coding [2] and sound synthesis, but has not been properly adopted in recognition and classification problems. A good review of the source-filter modeling work in instrument acoustics can be found in [3]. Here we do not stick to any exact physical interpretation of the model, but the goal is just to have a generic, compact, and accurate model for musical sounds.

For the sake of completeness, a frequency-dependent decay is included in the model. A method is proposed for estimating jointly the source, filter, and the decay parts. Pitch invariance is achieved by taking the fundamental frequency (F0) of the sounds into account explicitly: a guiding principle is that information about the instrument body response and other factors is obtained only at the positions of

the harmonic partials which “sample” the instrument timbre at the corresponding frequencies. Joint estimation of the source signal and the filter has been previously studied in speech processing, but typically with very different assumptions [4].

The proposed method is evaluated by measuring its accuracy in representing 33 different musical instruments. Also, we evaluate its usefulness for *auditory stream formation*, which refers to the task of separating sounds from a polyphonic signal and classifying consecutive sounds into streams associated with a certain sound source.

2. SIGNAL MODEL

An observed discrete-time signal $y(m) = s(m) + n(m)$ is assumed to consist of a clean sound $s(m)$ and additive i.i.d. Gaussian noise $n(m) \sim \mathcal{G}(n(m); 0, \sigma_n^2)$. In the frequency domain, this can be written as

$$Y_t(f) = S_t(f) + N_t(f), \quad (1)$$

where $Y_t(f)$ is the complex-valued Fourier spectrum of $y(m)$ in time frame t . For convenience, we describe the signal model using a continuous-valued frequency variable f .

The spectrum of $s(m)$ is further broken into its magnitude and phase parts, $S_t(f) = |S_t(f)|e^{i\angle S_t(f)}$. The magnitudes are modeled as

$$|S_t(f_h)| = \gamma X(h)B(f_h)L^t(f_h)E_t(f_h) \quad (2)$$

where $f_h \approx hF$, $h = 1, \dots, H$ is the frequency of the h^{th} overtone of a sound with fundamental frequency F .¹ Note that $|S_t(f)|$ is modeled only at the positions of the harmonics and is assumed zero elsewhere. That is, the model addresses only the periodic component of the sound. The scalar γ denotes the overall gain of the sound, $X(h)$ denotes levels of the harmonics at a vibrating source (“excitation”), $B(f_h)$ represents the frequency response of the instrument body (“filter”), and $L(f_h)$ is a loss filter which models the frequency-dependent decay of transiently-excited sounds and is near to zero for continuously-excited sounds. The frame counter t is reset to zero at the onset of each tone. $E_t(f_h)$ represents modeling error.

The phase spectrum $\angle S_t(f)$ of $s(m)$ is not modeled. This is because the phase relationships of different partials are often so irregular (varying from one tone to another) that it is difficult to learn a meaningful structure for them.

The problem addressed in this paper is to learn such $X(h)$, $B(f)$, and $L(f)$ that all sounds emitted by the instrument can be approximated using (2) with as little perceptual distortion as possible; minimizing $E_t(f_h)$ in a certain sense. In the following, we consider the spectrum $|S_t(f)|$ on a decibel scale. That is, we write (2) as

$$S_{\text{dB}}^{(t)}(f_h) = \gamma_{\text{dB}} + X_{\text{dB}}(h) + B_{\text{dB}}(f_h) + tL_{\text{dB}}(f_h) + E_{\text{dB}}^{(t)}(f_h) \quad (3)$$

¹The polyphonic case is discussed in Sect. 5.

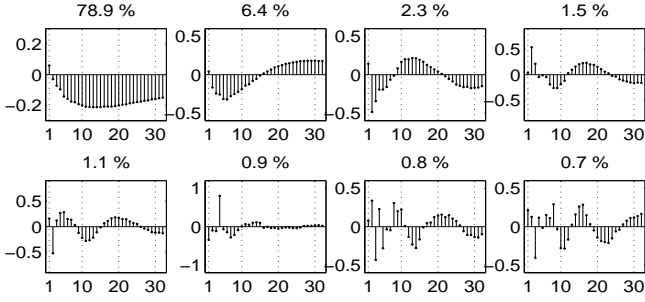


Fig. 1. Basis functions of $X_{dB}(h)$ found using PCA. The amount of data variance explained by each basis is shown on top of the panels.

where $S_{dB}^{(t)}(f) = 10 \log_{10}(|S_t(f)|^2)$, and similarly for the other terms. Avoiding the logarithm of zero is discussed later.

The harmonic levels $X_{dB}(h)$, body response $B_{dB}(f)$, and loss filter $L_{dB}(f)$ are further represented with a linear model so that the number of free parameters can be controlled:

$$X_{dB}(h) = \sum_{i=1}^{C_x} \xi_i x_i(h), \quad (4)$$

$$B_{dB}(f) = \sum_{j=1}^{C_b} \beta_j b_j(f) \quad \text{and} \quad L_{dB}(f) = \sum_{k=1}^{C_\ell} \lambda_k \ell_k(f).$$

This parametrises the model so that the $C_x + C_b + C_\ell$ parameters to be estimated (per instrument) are the weights ξ_i , β_j , and λ_k which define $X_{dB}(h)$, $B_{dB}(f)$ and $L_{dB}(f)$, respectively.

The basis functions $x_i(h)$, $b_j(f)$, and $\ell_k(f)$ can be chosen in many ways. In practice, it is useful to let the basis functions $b_i(f)$ and $\ell_i(f)$ to be identical and to choose them in a way that leads to a roughly constant resolution on a critical-band (CB) frequency scale. This is achieved by letting $b_i(f)$ consist of triangular band-pass responses that are distributed uniformly on a CB scale ($f_{CB} = 21.4 \log_{10}(0.00437f + 1)$) and overlap 50% with their neighbours.

Choice of the basis functions for $X_{dB}(h)$ is not so obvious. To find $x_i(h)$, we collected musical sounds from 33 musical instruments (as detailed in Sect. 6), measured the dB-levels of the first 32 harmonics in each sound and performed PCA for this data. Eliminating the effect of $B_{dB}(f)$ from the sounds prior to the PCA was found unnecessary. This is due to the fact that the effect of $B_{dB}(f)$ partly averages out when F0s of the sounds vary and the positions of the overtones move with respect to $B_{dB}(f)$.

Figure 1 shows the first eight basis functions $x_i(h)$ found using PCA. As can be seen, the principal components are close to the cosine basis but, interestingly, give a better resolution for the lower-order harmonics and tend to lump together higher-order partials.

3. OBSERVATION NOISE

This section describes how the levels $S_{dB}^{(t)}(f_h)$ of partials $h = 1, \dots, H$ in frames $t = 0, 1, \dots$ are extracted from noisy input signals $y(m)$. The data is then used to learn the instrument model later on. To estimate $S_{dB}^{(t)}(f_h)$, we need to analyze the statistics of the power spectrum $|Y_t(k)|^2$ which consists of the clean sound and noise. Note that now we consider discrete spectra with index k . The spectrum $Y_t(k)$ is calculated by Hamming-windowing a time frame, zero-padding it to four times its length, and by applying the Fourier transform.

Given that the noise $n(m)$ is Gaussian in the time domain, the real and imaginary part of the Fourier coefficients $N_t(k)$ are also

Gaussian and have variance $\sigma_N^2 = dM\sigma_n^2/2$, where σ_n^2 is noise variance in the time domain, M is the time frame length in samples, and $d \approx 0.4$ is the mean of a squared Hamming window. The magnitude spectrum values $|N_t(k)|$ are Rayleigh distributed, $|N_t(k)| \sim \mathcal{R}(x; \sigma_N^2) = \frac{x}{\sigma_N^2} \exp\left(-\frac{x^2}{2\sigma_N^2}\right)$, and the phase spectrum $\angle N_t(k)$ is uniformly distributed, $\angle N_t(k) \sim \mathcal{U}_{[0, 2\pi]}(\cdot)$ [5, p.160].

For convenience, let us omit both time and frequency indices for a while and consider an individual frequency component $Y_{t'}(k')$ at an arbitrary frame t' and frequency k' . As a shorthand, we denote $Y = Y_{t'}(k')$, $a_S = |S_{t'}(k')|$, $a_N = |N_{t'}(k')|$ and the phase difference between the clean sound and noise as $\theta = \angle S_{t'}(k') - \angle N_{t'}(k')$. Due to the noise characteristics, θ is uniformly distributed.

The power spectrum value $|Y|^2$ can be written as

$$|Y|^2 = |a_S + a_N e^{i\theta}|^2 = a_S^2 + 2a_S a_N \cos \theta + a_N^2. \quad (5)$$

Expectation of $|Y|^2$ can be calculated as

$$\mu_{Y^2} = \int_0^\infty \int_{-2\pi}^{2\pi} |Y|^2 p(\theta) p(a_N) d\theta da_N = a_S^2 + 2\sigma_N^2, \quad (6)$$

where we have used the identity $\mathbb{E}_{p(x)}(x^{2n}) = 2^n n! \sigma_N^{2n}$ for even moments of the Rayleigh density.

An estimate of $S_{dB}^{(t)}(k)$ is then obtained by

$$\hat{S}_{dB}^{(t)}(k) = 10 \log_{10} (g(|Y_t(k)|^2)), \quad (7)$$

where $g(x) = \max(x - 2\sigma_N^2, \epsilon)$ removes the bias caused by additive noise and constraints resulting negative or near-zero values to a small constant $\epsilon = 10^{-2} \cdot 2\sigma_N^2$ which prevents logarithm of zero.

The level $S_{dB}^{(t)}(f_h)$ of a certain partial h is estimated using the highest local maximum in $\hat{S}_{dB}^{(t)}(k)$ around the frequency of the partial. For isolated tones, the search range is $[(h - \frac{1}{2})F, (h + \frac{1}{2})F]$ whereas for polyphonic signals this has to be narrower.

Next, let us consider the variance $\sigma_{\hat{S}_{dB}}^2$ of an estimated partial level. This depends on the level itself, variance being lower for high-level partials. The variances are important in order to weight different observations according to their reliability. The variance of $|Y|^2$ is given by

$$\begin{aligned} \sigma_{Y^2}^2 &= \int_{a_N} \int_{\theta} (|Y|^2 - a_S^2 - 2\sigma_N^2)^2 p(\theta) p(a_N) d\theta da_N \quad (8) \\ &= 4\sigma_N^2 (a_S^2 + \sigma_N^2). \end{aligned}$$

The variance $\sigma_{\hat{S}_{dB}}^2$ can then be calculated using a numerical technique called unscented transform [6]. The formulae can be found in the reference, but the basic idea is to estimate how the variance of $|Y|^2$ changes in the non-linear $\log(\cdot)$ function in (7). It was found that the variance $\sigma_{\hat{S}_{dB}}^2$ is inversely proportional to a linear function of $|Y|^2$.

4. PARAMETER LEARNING

Now the actual goal is to learn the instrument model parameters ξ_i , β_j , and λ_j , when given observed levels of harmonics $\hat{S}_{dB}^{(t)}(f_h)$, $h = 1, \dots, H$, in analysis frames that cover various sounds produced by the instrument in question.

The observed levels $\hat{S}_{dB}^{(t)}(f_h)$ are used to build a system of equations from which the model parameters are estimated. Since the absolute gain γ_{dB} of the analysed sounds is not of interest, we consider

the partial levels only in relation to each other. Using (3), the information provided by each pair of observed harmonics h and j in frame t can be written as

$$\begin{aligned} \hat{S}_{\text{dB}}^{(t)}(f_h) - \hat{S}_{\text{dB}}^{(t)}(f_j) & \\ = \gamma_{\text{dB}} + X_{\text{dB}}(h) + B_{\text{dB}}(f_h) + tL_{\text{dB}}(f_h) + E_{\text{dB}}^{(t)}(f_h) & \\ - \gamma_{\text{dB}} - X_{\text{dB}}(j) - B_{\text{dB}}(f_j) - tL_{\text{dB}}(f_j) - E_{\text{dB}}^{(t)}(f_j). & \end{aligned} \quad (9)$$

Substituting the basis-function representations (4), this becomes

$$\begin{aligned} \sum_{i=1}^{C_x} \xi_i [x_i(h) - x_i(j)] + \sum_{j=1}^{C_b} \beta_j [b_j(f_h) - b_j(f_j)] & \\ + t \sum_{k=1}^{C_\ell} \lambda_k [\ell_k(f_h) - \ell_k(f_j)] = \hat{S}_{\text{dB}}^{(t)}(f_h) - \hat{S}_{\text{dB}}^{(t)}(f_j) + E_{\text{dB}}^*. & \end{aligned} \quad (10)$$

where $E_{\text{dB}}^* = E_{\text{dB}}^{(t)}(f_j) - E_{\text{dB}}^{(t)}(f_h)$ is treated as a random variable.

The above can be written in a matrix form as

$$\mathbf{Q}\mathbf{u} = \mathbf{a} + \mathbf{e}, \quad (11)$$

where the vector $\mathbf{u} = [\xi_1, \dots, \xi_{C_x}, \beta_1, \dots, \beta_{C_b}, \lambda_1, \dots, \lambda_{C_\ell}]^T$ contains the parameters to be learned, and each pair of harmonics generates one row to the matrix \mathbf{Q} , the row being $[x_1(h) - x_1(j), x_2(h) - x_2(j), \dots, b_1(f_h) - b_1(f_j), b_2(f_h) - b_2(f_j), \dots, t[\ell_1(f_h) - \ell_1(f_j)], t[\ell_2(f_h) - \ell_2(f_j)], \dots]$. The vector \mathbf{a} contains the observed decibel level differences $\hat{S}_{\text{dB}}(f_h) - \hat{S}_{\text{dB}}(f_j)$ corresponding to each row of \mathbf{Q} , and \mathbf{e} contains the terms E_{dB}^* . Provided that H partials are observed in frame t , exactly $H - 1$ linearly independent equations (rows of \mathbf{Q}) can be generated. We do this by pairing the strongest partial with all the remaining $H - 1$ partials. Note that H varies from sound to sound depending on their F0s, since we can use only partials that fall within the frequency range of the body response.

Similarly to the above equations, each pair of harmonics in two consecutive frames t and $t - 1$ of a same sound provide information about the unknown parameters. If H partials are observed in the two frames, H linearly independent equations can be generated. Without loss of generality, we can pair partials with the same harmonic index h in the two frames. The resulting equation is

$$\begin{aligned} \hat{S}_{\text{dB}}^{(t)}(f_h) - \hat{S}_{\text{dB}}^{(t-1)}(f_h) + E_{\text{dB}}^* & \\ = B_{\text{dB}}(f_h^{(t)}) + tL_{\text{dB}}(f_h^{(t)}) - B_{\text{dB}}(f_h^{(t-1)}) - (t-1)L_{\text{dB}}(f_h^{(t-1)}), & \end{aligned} \quad (12)$$

where $E_{\text{dB}}^* = E_{\text{dB}}^{(t)}(f_h) - E_{\text{dB}}^{(t-1)}(f_h)$ is treated as a random variable and the terms γ_{dB} and $X_{\text{dB}}(h)$ do not appear because they cancel out. The time-dependence of f_h is underlined by writing $f_h^{(t)}$, since the F0 may vary from frame to frame even within a single sound.

By substituting the basis function representations (4) to (12) and writing the result in a matrix form, the following row is generated to \mathbf{Q} : $[0, \dots, 0, b_1(f_h^{(t)}) - b_1(f_h^{(t-1)}), b_2(f_h^{(t)}) - b_2(f_h^{(t-1)}), \dots, t\ell_1(f_h^{(t)}) - (t-1)\ell_1(f_h^{(t-1)}), t\ell_2(f_h^{(t)}) - (t-1)\ell_2(f_h^{(t-1)}), \dots]$. The corresponding level differences $\hat{S}_{\text{dB}}^{(t)}(f_h) - \hat{S}_{\text{dB}}^{(t-1)}(f_h)$ are stored in \mathbf{a} , and E_{dB}^* goes into \mathbf{e} .

The error terms E_{dB}^* and E_{dB}^* represent all the intrinsic variation in the played sounds that cannot be captured by the source-filter-decay model. These include factors such as excitation dynamics, plucking point variation, etc. Together with the additive noise, they affect the variance of the level differences $\hat{S}_{\text{dB}}(\cdot) - \hat{S}_{\text{dB}}(\cdot)$ stored in \mathbf{a} . We use a common std value $\sigma_E = 1$ [dB] for both E_{dB}^* and E_{dB}^* . This is slightly optimistic: values around 1–6 dB are reasonable.

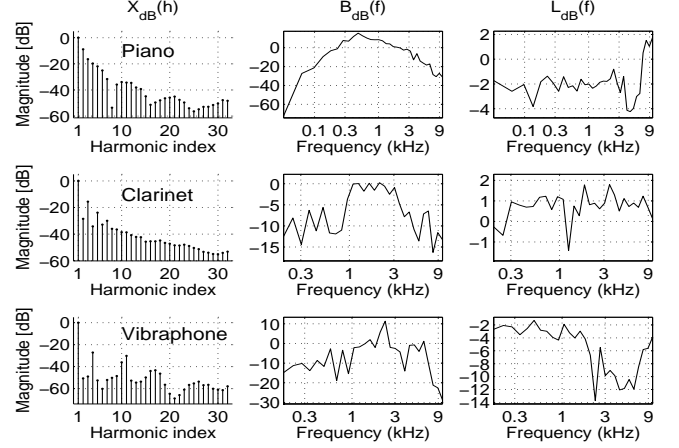


Fig. 2. Functions $X_{\text{dB}}(h)$, $B_{\text{dB}}(f)$, and $L_{\text{dB}}(f)$ learned for three instruments: piano (top), clarinet (middle), and vibraphone (bottom).

As the absolute level of the sounds is not fixed, it is necessary to normalise ξ_i and β_j to make them well-defined. This can be done for example by requiring that the level of $X_{\text{dB}}(h)$ for the first harmonic is 0 dB by writing $\sum_{i=1}^{C_x} \xi_i x_i(1) = 0$, and that the level of the body response at 1000 Hz is 0 dB by writing $\sum_{j=1}^{C_b} \beta_j b_j(1000) = 0$. The corresponding rows are then added to \mathbf{Q} and \mathbf{a} .

There are generally more observations than unknowns, and therefore (11) is overdetermined. A solution which minimizes the least-square (LS) error criterion $\|\mathbf{Q}\mathbf{u} - \mathbf{a}\|^2 = \|\mathbf{e}\|^2$ is readily obtained by the LS estimator $\hat{\mathbf{u}} = (\mathbf{Q}^T \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{a}$. An extension of this is the weighted LS estimator

$$\hat{\mathbf{u}} = (\mathbf{Q}^T \mathbf{W} \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{W} \mathbf{a}, \quad (13)$$

where a natural choice for \mathbf{W} is a diagonal matrix with weights $[\mathbf{W}]_{i,i}$ inversely proportional to the variance of each observed level difference in the vector \mathbf{a} . More exactly,

$$[\mathbf{W}]_{i,i} = 1/(\sigma_{\hat{S}_{\text{dB}}}^2(i, 1) + \sigma_{\hat{S}_{\text{dB}}}^2(i, 2) + \sigma_E^2), \quad (14)$$

where $\sigma_{\hat{S}_{\text{dB}}}^2(i, 1)$ and $\sigma_{\hat{S}_{\text{dB}}}^2(i, 2)$ are the additive noise variances of the two partials whose level difference was stored in element i of \mathbf{a} , and σ_E^2 is as described above. In practice, the variances $\sigma_{\hat{S}_{\text{dB}}}^2$ are negligible for all except the weakest partials.

Figure 2 shows examples of learned $X_{\text{dB}}(h)$, $B_{\text{dB}}(h)$, and $L_{\text{dB}}(h)$ for three different instrument. One-second long sounds covering the entire pitch range of each instrument were used to estimate the parameters. For piano (top panels), we can see that every 8th harmonic is missing from $X_{\text{dB}}(h)$ because the strings are excited at 1/8 point along their length. For piano, $B_{\text{dB}}(h)$ does not model only the resonances of the soundboard, but also the varying number of significant partials in the lowest- (27Hz) and the highest-pitched (4.2kHz) tones. Combining $X_{\text{dB}}(h)$ with $B_{\text{dB}}(h)$ is successful in representing both extremes. Loss filter shows slight attenuation for all except the highest partials, where some energy appears to be transferred from the lower partials. Characteristics of the clarinet and the vibraphone can be observed in the middle and the lower panels.

5. AUDITORY STREAM FORMATION

Here we attempt to analyze a polyphonic music signal so as to estimate the F0s and chain together sounds that belong to a same instrument. The LS estimation framework is particularly suitable for

this task, because it allows sequential computation. In the sequential LS, the parameter vector $\hat{\mathbf{u}}$ is initialised to all-zero and then updated sequentially, using one row at a time from \mathbf{Q} and \mathbf{a} . Also the corresponding LS error, $J = \|\mathbf{Q}\mathbf{u} - \mathbf{a}\|^2$, can be calculated sequentially. The update equations can be found in [7, p.249].

We perform auditory streaming in a Viterbi-like manner as follows: First, a polyphonic input signal is processed with an instrument-independent multiple-F0 estimator which finds R F0s at each temporal segment q of the input. In segment 0, R different sequential-LS streams are initialised, each corresponding to a unique instrument model. In segment q , each of the $r = 1, \dots, R$ tones detected in the segment are taken into consideration one at a time. The sound r is appended on trial to all the R streams ending at the preceding segment $q - 1$, and the one leading to the smallest LS error J is chosen to be the stream for that sound. After all sounds in all frames are processed, the best stream is found by backtracking from the sound with the smallest LS error in the last segment Q .

The described method requires that F0s of the sounds and their partial levels can be estimated from the mixture signal. We used the multiple-F0 estimator [8] for this purpose, and estimated partials levels directly by picking spectral maxima nearby their frequencies.

6. RESULTS

The proposed method was evaluated by training models for 33 different musical instruments. The data consisted of the McGill University Master Samples collection, independent recordings for the acoustic guitar, and Roland XP30 samples for the hammond organ and the electric guitar. Sounds over the entire pitch range of each instrument were used and partial levels were estimated in 93 ms frames over the leading one second of each sound.

The model accuracy was evaluated by measuring average perceptual distortion between the model output and the original samples. Perceptual distortion was computed by comparing the dB levels at each critical band for the model and for the input data. The dB-level differences were squared, averaged over the active bands of a sound, and the resulting values were averaged over different sounds of the same instrument. These values were averaged over instruments, and at the end, a square root was taken. The absolute levels of the model and the data were matched in each individual analysis frame; that is, only the shape of the spectrum was of interest.

Figure 3 shows the results for four different model configurations. The model XBL refers to the source-filter-decay model where $X_{dB}(\cdot)$, $B_{dB}(\cdot)$, and $L_{dB}(\cdot)$ were all used, the model BL refers to a configuration where only $B_{dB}(\cdot)$ and $L_{dB}(\cdot)$ were included in the linear model (3), and so forth. The model B represents the traditional MFCC model, since the (linear) cosine transform would not make any difference to the model accuracy. As can be seen, the model X alone is not sufficient for any instrument family. The model B (cf. MFCCs) and BL perform surprisingly well for many instrument families, but only the model XBL performs well for all. Sounds that benefit significantly from the term $X_{dB}(\cdot)$ include the mallet percussions, clarinets, hammond organ, and the electric guitar.

Auditory streaming was evaluated as follows. Melodic lines were extracted from MIDI files and synthesised with samples from above-described instrument database, however simplifying the time structure so that all notes were 280 ms in length. Two melodies were then randomly mixed, but ensuring that no identical F0s were played simultaneously in the two melodies, and that the two instruments belonged to different instrument families (listed in Fig. 3). The method described in Sect. 5 was used to estimate the F0s and to chain together notes coming from a same instruments. 48 mixture signals

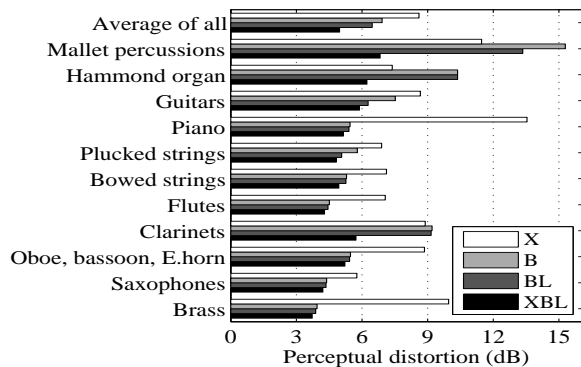


Fig. 3. Perceptual distortion caused by different models, averaged over all instruments (top) and within instrument families.

were generated, each consisting of 28 notes on the average. The proposed method was able to find the F0s and stream correctly 88% of the notes. In higher polyphonies, the sound separation part should be done more carefully than just picking partials from the spectrum.

7. CONCLUSIONS

A new technique for modeling musical instrument sounds was proposed. The simulation results show clear improvement over MFCC-like models for certain instrument families. Better modeling accuracy is expected to lead to better classification and coding results in the future work. The proposed method for auditory streaming was shown to perform well for low-polyphony material where the employed sound separation technique was sufficient.

8. ACKNOWLEDGEMENTS

Thanks to Taylan Cemgil from Cambridge University for fruitful discussions regarding the observation noise model. This work was supported by the Academy of Finland, project 213462.

9. REFERENCES

- [1] N.H. Fletcher and T.D. Rossing, *The Physics of Musical Instruments*, Springer, Berlin, Germany, second edition, 1998.
- [2] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *IEEE ICASSP*, Tampa, Florida, 1985, pp. 937–940.
- [3] V. Välimäki, J. Pakarinen, C. Erku, and M. Karjalainen, "Discrete-time modelling of musical instruments," *Reports on progress in physics*, vol. 69, pp. 1–78, 2006.
- [4] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 2, pp. 492–501, 2006.
- [5] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, IEEE Press, New York, 1987.
- [6] S.J. Julier and J.K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Tech. Rep., Eng. Dept., Oxford University, 1994.
- [7] Steven M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.
- [8] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006.