44º CONGRESO ESPAÑOL DE ACÚSTICA
ENCUENTRO IBÉRICO DE ACÚSTICA
EAA EUROPEAN SYMPOSIUM ON ENVIRONMENTAL
ACOUSTICS AND NOISE MAPPING

INSTRUMENT MODELS AND ITS APPLICATIONS

PACS: 43.60.-C.

Pedro Vera-Candeas; Francisco J. Cañadas-Quesada; Pablo Cabañas-Molero; Francisco J. Rodríguez Serrano.
Universidad de Jaén.
Calle Alfonso X el Sabio, 28. E.P.S. Linares Linares , Jaén. España.
Tel: 953 648 554.
E-mail: pvera@ujaen.es

ABSTRACT

A musical instrument can be recognized by its unique timbre. One way to parameterize the timbre of an instrument is to obtain the spectrum for each musical note. The use of instrument models makes possible to approximately estimate the spectrum produced by a harmonic instrument for each musical note. This information can be incorporated to improve a bunch of musical processing applications: score following, audio restoration, source separation or transcription per instrument. In this work, we show the results obtained when instrument models are used in these applications.

RESUMEN

El timbre es una de las características intrínsecas a los instrumentos musicales. Uno de los parámetros que define el timbre de un instrumento es el espectro que produce para cada nota musical. Mediante el uso de modelos de instrumento es posible realizar una estimación aproximada del espectro que se obtiene para cada nota musical de un instrumento armónico. Esta información se ha utilizado recientemente en la mejora de una serie de aplicaciones de interés en el campo del procesado de música: alineamiento música-partitura, restauración de audio, separación de instrumentos o la transcripción por instrumento. En este trabajo se presentan los resultados de estas aplicaciones cuando se usa la información de modelos de instrumento.

1. INTRODUCTION

Approaches that model an audio spectrogram as a linear combination of sound objects have been recently successfully used in applications such as sound source separation [1], melody extraction [2], music transcription [3] and sound source recognition [4]. In this context, the short-term magnitude (or power) spectrum of the signal *x(f,t)* in frame *t* and frequency *f* is modeled as a weighted sum of basis functions as

$$\hat{x}_t(f) = \sum_{n=1}^{N} g_{n,t} b_n(f) \qquad (1)$$

where $g_n(t)$ is the gain of the basis function n at frame t, and $b_n(f)$, $n = 1,...,N$ are the bases. When dealing with harmonic instruments sounds in the context of automatic music transcription, each basis function ideally represents a single pitch, and the corresponding gains contain information about the onset and offset times of notes having that pitch.

2. INSTRUMENT MODELS

2.1. Excitation-Filter Model

The main problem of the model presented in Eq. (1) is that it requires a distinct basis function to represent each pitch of each instrument. Thus, a large number of parameters that are not tied between different pitches has to be tuned, making difficult to estimate or adapt the model. To reduce the complexity, Virtanen and Klapuri [5] proposed to model each basis as the product of the magnitude spectra of an excitation $e_n(f)$ and a filter $h_j(f)$. Each basis function is indexed by excitation n and filter j:

$$b_{n,j}(f) = h_j(f)e_n(f), \quad n = 1, ..., N, j = 1, ..., J, \tag{2}$$

where N is the number of excitations and J the number of filters. Typically each instrument is represented using a single filter that corresponds to the resonant structure of the body of the instrument. This significantly reduces the number of parameters. However, since a piece of music can contain many different pitches and for each pitch a full spectrum is needed to represent $e_n(f)$, there are still many parameters to tune.

The excitation-filter (or source-filter) model has origins in speech processing and sound synthesis. In speech processing the excitation models the sound produced by the vocals cords, whereas the filter models the resonating effect of the vocal tract. The voiced part of the speech excitation can be modeled as a train of pulses, which results in a harmonic excitation, where the amplitudes of the harmonics are smooth as a function of frequency. In sound synthesis, excitation-filter synthesis colors a spectrally rich excitation signal to get the desired sound.

2.2. Harmonic Comb Excitation

Another way to restrict the model in Eq. (1) deals with the harmonicity. Musical notes, excluding transients, are pseudo- periodic, and their spectra consists of regularly spaced frequency peaks. Therefore, we assume that the elements in the basis $b_n(f)$ or $b_{n,j}(f)$ should follow this harmonic shape.

Several studies ([4], [1]) consider the excitation as frequency components of unity magnitude at integer multiples of a certain fundamental frequency. This results in modeling the excitation using a harmonic comb consisting of a sum of harmonic components as:

$$e_n(f) = \sum_{m=1}^{M} G(f - mf_0(n)) \tag{3}$$

where $m = 1, ..., M$ is the number of harmonics, $f_0(n)$ the fundamental frequency of excitation n. $G(f)$ is the magnitude spectrum of the window function, and the spectrum of a harmonic component at frequency $mf_0(n)$ is approximated by translated $G(f - mf_0(n))$.

The harmonic comb excitation is not suitable for certain types of instruments that do not possess a smooth nature in frequency.

2.3. Harmonic Multi-Excitation Model

An interesting alternative is the use of the multi-excitation model proposed in [6]. This model defines the excitation spectrum as a linear combination of a few excitation basis vectors. The dimensions of the vectors are harmonic indices, i.e. the first dimension corresponds to the first harmonic and the second dimension corresponds to the second harmonic, etc. The excitation vectors are instrument-dependent but are not pitch-dependent. This model is going to be explained in detail in this section.

A generic excitation model represents the spectral basis functions as

$$b_{n,j}(f) = h_j(f) e_{n,j}(f). \tag{4}$$

A generic harmonic excitation is defined as

$$e_{n,j}(f) = \sum_{m=1}^{M} a_{m,n,j} G\left(f - m f_0(n)\right) \tag{5}$$

where $a_{m,n,j}$ is the amplitude of partial (or harmonic) m, pitch n and instrument j. We propose to model the amplitudes as the linear combination of I excitation basis vectors $v_{i,m,j}$ as

$$a_{m,n,j} = \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} \tag{6}$$

where $w_{i,n,j}$ is the weight of the i-th excitation basis vector for pitch n and instrument j. The excitation bases are unique for each instrument and harmonic but shared across pitches, whereas the weights are unique for each instrument and pitch, but shared between harmonics. Substituting equation (6) into (5), the harmonic excitation functions can be expressed as

$$e_{n,j}(f) = \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G\left(f - m f_0(n)\right) \tag{7}$$

These harmonic excitation functions are multiplied by the instrument filter to obtain the spectral basis functions as expressed in equation (4). Finally, the model for magnitude spectrum of a whole signal frame is the sum of instruments and pitches given as

$$\hat{x}_t(f) = \sum_{n,j} g_{n,t,j} h_j(f) \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G\left(f - m f_0(n)\right) \tag{8}$$

where $n = 1,...,N$ (N being the number of pitches) and $j = 1, ..., J$ (J being the number of instruments). M represents the number of harmonics and I the number of considered excitations with $I << N$.

Using a small number of excitation bases reduces significantly the parameters of the model, which benefits to the learning of parameters. In this model, $g_{n,t,j}$ represents the gains applied to pitch n for instrument j at frame t. $v_{i,m,j}$ is the m-th partial of the i-th excitation basis for instrument j. $w_{i,n,j}$ are the excitation weights, that is, the weights indicate the proportion of i-th excitation basis for each pitch n and instrument j. Finally, $h_j(f)$ represents the instrument filter.

Non-negativity of the parameters has turned out to be an efficient constraint in learning the spectrogram factorization models [5], when dealing with amplitudes is also a natural restriction. Thus, we restrict all the parameters of the model (8) to non-negative values (Non-negative Matrix Factorization, NMF, is usually called this framework). Under these restrictions, we estimate the parameters by minimizing the reconstruction error between the observed spectrogram and the model one. In order to obtain the values of model parameters that minimize the cost function, [7] proposes an iterative algorithm based on multiplicative update rules. Further details can be found in [7].

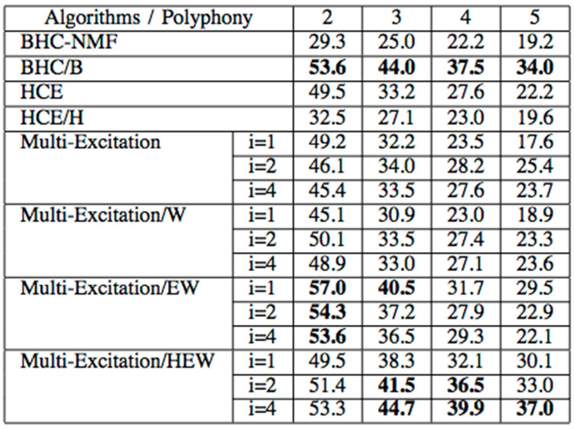
3. APPLICATIONS

3.1. Instrument-Specific Transcription.

The instrument models are able to discriminate the instrument j from the detected notes because these models estimate the gains $g_{n,j}(t)$ as a function of the instrument. Using this information, it is possible to produce instrument-specific transcription. The average results for the polyphonic database proposed in [3] are given in Table I. Here we consider a note to be correct only when is produced by the correct instrument.

We do NMF-based transcription using three different models. 1) The Basic harmonic constrained NMF (BHC-NMF) model, which extends the model proposed in Eq. (1) by introducing a harmonicity constraint. 2) The Harmonic Comb Excitation (HCE) model defined in Eq. (3). 3) The Multi-Excitation (MEI) model proposed in Eq. (7).

No comparison to previous work has been made in this task because all the state-of-the-art methods are unsupervised, and do not produce instrument-specific transcription. Logically, without knowing the instruments in advance it is not possible to classify the notes between the different sources. The 95% confidence intervals of the average F-measure (F) are 3.7% for all the models. The numerical results in bold presented in Table I are the best and the statistically similar to them.



Algorithms / Polyphony		2	3	4	5
BHC-NMF		29.3	25.0	22.2	19.2
BHC/B		**53.6**	**44.0**	**37.5**	**34.0**
HCE		49.5	33.2	27.6	22.2
HCE/H		32.5	27.1	23.0	19.6
Multi-Excitation	i=1	49.2	32.2	23.5	17.6
	i=2	46.1	34.0	28.2	25.4
	i=4	45.4	33.5	27.6	23.7
Multi-Excitation/W	i=1	45.1	30.9	23.0	18.9
	i=2	50.1	33.5	27.4	23.3
	i=4	48.9	33.0	27.1	23.6
Multi-Excitation/EW	i=1	**57.0**	**40.5**	31.7	29.5
	i=2	**54.3**	37.2	27.9	22.9
	i=4	**53.6**	36.5	29.3	22.1
Multi-Excitation/HEW	i=1	49.5	38.3	32.1	30.1
	i=2	51.4	**41.5**	**36.5**	33.0
	i=4	53.3	**44.7**	**39.9**	**37.0**

Table I. Average F-Measure on polyphonic woodwind data per instrument.

The best transcription results are obtained by the Multi-Excitation and BHC-NMF models. Possible causes of the HCE model underperformance is due to the inaccurate modeling for some instruments like the clarinet. Besides, for the Multi-Excitation model, the results sometimes increase when more excitations are considered.

The model parameters can be adapted to the test signals in order to update to instrument model to the test databases (the model parameters are tuned to the training database which is composed of isolated sounds from the RWC database [8]). In Table I, the parameters labelled with "/" are not updated at the test stage. Adaptation of the model parameters does not improve the results except for the HCE model and the Multi-Excitation model except for polyphony 2. As we could see in the previous experiments, adaptation suffers from the huge number of free parameters. This would be the reason of the better performance when adapting the filter for the HCE model.

Finally, as expected, the transcription results per instrument decrease with the level of polyphony. This means that more notes are attributed to the wrong instrument when the level of polyphony increases.

3.2. Source Separation

In this application, we deal with the problem of online separation of harmonic musical sources from a single-channel recording. A score-informed SSS system with instrument models is implemented. It uses initial instrument models and score information as prior information, the instrument models are updated with the aid of the aligned score towards the real played instrument. All the system modules can also be run in an offline fashion depending on the concrete application.

The proposed system needs the aligned score information. Because of that, an alignment stage is implemented as described in [9]. This alignment information is used in an online NMF framework to compute the factorization and the online instrument model updating. The NMF framework is used as a signal decomposition method as in [6] and it is initialized with the score information and instruments models that represent the spectral shape of each instrument (using the MEI model of Eq (8)). Here, we use the NMF framework of [6], which is adapted in order to run online [10].

We compare different configurations for proposed method to a baseline score-informed source separation methods proposed in [9], denoted as Soundprism. It separates sources using harmonic masking where the energy of overlapping harmonics are distributed according to the harmonic indices of the sources. It is an online algorithm but no instrument models are used. The source separation method has three configurations. Proposed fixed denotes the offline version of the proposed method using fixed instrument models. Proposed adaptive offline denotes the offline version of the proposed method with adaptive instrument models, and Proposed adaptive online denotes its online version. We also compare with Oracle, the theoretically best source separation method based on time-frequency masking methods and the analysis filter bank used on the source separation system. Its calculation requires the isolated sound sources.

We compare source separation methods taking audio-score alignment results (i.e. the refined score pitches) as inputs. This gives us the realistic results. Figure 1 shows the results on recordings of different polyphonies, SDR values are shown. The average SDR of all methods except Oracle degrades and the standard deviation increases. This is intuitive, as the audio-score alignment errors are responsible for these degradations. Second, with the increase of polyphony, the degradations are less significant for almost all methods. This can be explained by the performance of the audio-score alignment. On the dataset proposed in [9], the alignment was better on pieces with higher polyphony. Third, for all polyphonies, the baseline method, Soundprism, degrades most significantly, while the degradations of the proposed method are much less. This causes the performance gap between Soundprism and the proposed method even larger when working with audio-score alignment. This is promising, as it indicates the advantage of instrument models in realistic score-informed source separation scenarios.

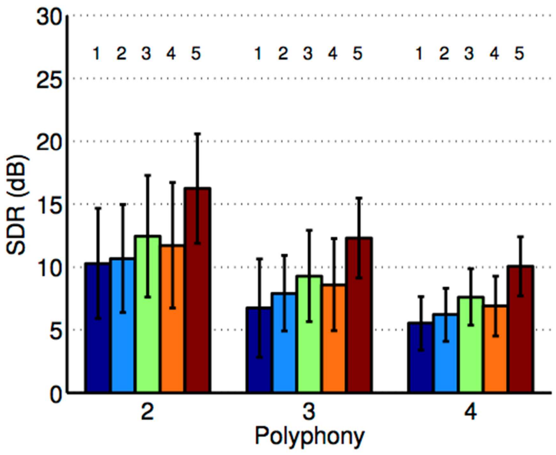


Figure 1. Average and standard deviation of source separation results versus polyphony, calculated using the alignment information. The five methods are 1) Soundprism, 2) Proposed fixed, 3) Proposed adaptive offline, (4) Proposed adaptive online, 5) Oracle.

3.3. Audio Restoration

In this section, we present a constrained non-negative matrix factorization approach to isolate the target source (piano) from a piano signal degraded by vinyl noise. The audio restoration framework is trained with spectral patterns for piano sounds (using the MEI model) and vinyl noise. Results show that the use of instrument models and sparsity constraints improves the separation capabilities in terms of Signal-to-Distortion-Ratio (SDR) in comparison with some commercial software.

Two constraints motivated by the sparsity principle are here utilized: monophony and polyphony [11]. The first one is designed for activating just one basis at each frame of the factorization. In order to achieve this goal, the cross correlations between the components of the gain matrix $g_{n,j}(t)$ are added as a reguralization term to the global distortion. The second one, polyphony, is an evolution of the monophony. In this case, the information of the score of the piano excerpt is required. The score is characterized for activating at the same time a set of notes that represent the different states of the score. This constraint allows the activation at the same frame of those combinations of notes presented in the score. On the contrary, concurrent activations of notes that do not occur in the score are penalized by a reguralization term based on cross correlations between spectral patterns. More details can be obtained in [11].

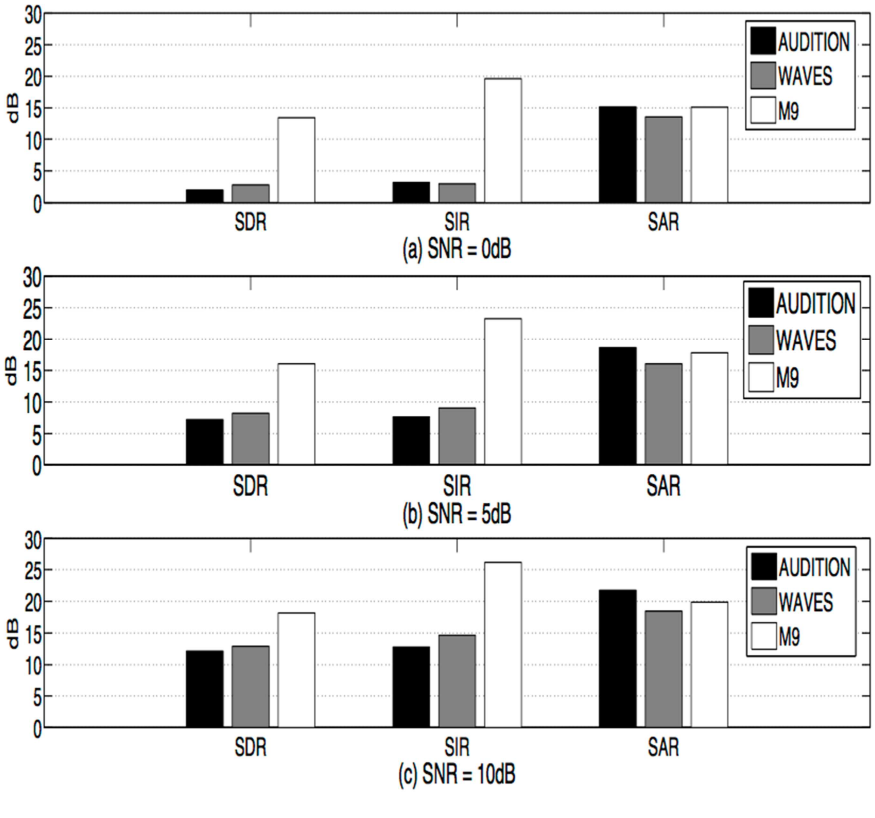


Figure 2. SDR, SIR and SAR piano results comparing an instrument model based method with sparsity constraints (M9) and two commercial audio restoration softwares (AUDITION and WAVES) evaluating at Signal-to-Noise Ratio of (a) 0 dB, (b) 5 dB and (c) 10 dB.

A comparison between an instrument model based method (M9) and two commercial audio restoration software (AUDITION and WAVES) are presented in Figure 2. The used database is obtained mixing clean piano excerpts from MAPS database with recorded vinyl noise at different Signal-to-Noise Ratios. Figure 2(a) shows that the best SDR and SIR results are provided by our method (using monophonic and polyphonic constraints for training and testing). It can be seen that SDR and SIR results provided by our method improves, about 10dB and 16dB, SDR and SIR results from both commercial software. However, SAR results are similar in all of them. Figure 2(b) shows that the best SDR and SIR results are still provided by our method but now, the SDR and SIR improvement is reduced to 8dB and 14dB. In a similar way, the best SDR and SIR results are achieved by our method in Figure 2 (c) but the improvement is about 6dB and 12dB in terms of SDR and SIR results. Comparing Figure 2(a), Figure 2(b) and Figure 12(c) , it can be observed that these commercial audio restoration software does not work adequately evaluating a low SNR. Specifically, these methods fail when the noise profile exhibit fast changes along the time.

In order to give the reader the opportunity of listening the material, a webpage for the results has been created. On this page, some audio examples (degraded, separated piano and separated noise) can be heard by the reader. The web page can be found at https://dl.dropboxusercontent.com/u/22448214/JNMRSIMML2013/index.html

4. CONCLUSIONS

The use of instrument models helps to describe each source and this information is very useful for a factorization framework, as NMF, which tries to identify some patterns at the mixed signal. The most reliable is the instrument model, the better factorization the NMF does, and so, the better results are obtained. This aim of using the best instrument model as possible motivates to update the initial models up to the real played instrument patterns. The possibilities of updating the models with information from the mixed signal is reduced by the corruption of the spectral information when the instruments are playing at the same time. However, it is important to take advantage of all the available information at the mixed signal in order to improve the model, because until the most little details obtained from the signal that is going to be factorized can be decisive.

REFERENCES

[1] A. Klapuri, T. Virtanen, T. Heittola, “Sound source separation in monaural music signals using excitation-filter model and em algorithm,” in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Dallas, USA, 2010.

[2] J.L. Durrieu, G. Richard, B. David, C. Fevotte, “Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Sig- nals,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18 , no. 3, pp. 564 - 575, March 2010.

[3] E. Vincent, N. Bertin, 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, March 2010.

[4] T. Heittola, A. Klapuri, T. Virtanen, “Musical instrument recognition in polyphonic audio using source-filter model for sound separation,” Proc. 10th Int. Society for Music Information Retrieval Conf. (ISMIR), Kobe, Japan, 2009.

[5] Virtanen, T., Klapuri, A., “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, 2006.

[6] Carabias-Orti, J.J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., and Cañadas-Quesada, F.J. (2011). “Musical Instrument Sound Multi-Excitation Model for Non-Negative Spectrogram Factorization”. IEEE Journal of Selected Topics in Signal Processing, Vol. 5, no. 6, pp. 1144-1158.

[7] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. of Neural Information Processing Systems, Denver, USA, 2000.

[8] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, “RWC Music Database: Popular, Classical, and Jazz Music Databases,” Proc. of the 3rd Int. Society for Music Information Retrieval Conf. (ISMIR), Paris, France, October 2002.

[9] Duan, Z., and Pardo, B. (2011). “Soundprism: An Online System for Score-Informed Source Sep- aration of Music Audio”. Selected Topics in Signal Processing, IEEE Journal of, vol.5, no.6, pp.1205-1215, doi: 10.1109/JSTSP.2011.2159701.

[10] Lefevre, A., Bach, F., and Fevotte, C. (2011). “Online algorithms for Nonnegative Matrix Factorization with the Itakura-Saito divergence”. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011.

[11] Cañadas-Quesada, F.J., Vera-Candeas, P. and Ruiz-Reyes, N. “Monophonic/Polyphonic Constrained Non-Negative Matrix Factorization Applied to Piano Restoration”. Journal of New Music Research, Submitted.