Unsupervised Single-channel Music Source Separation by Average Harmonic Structure Modeling

Zhiyao Duan*, Yungang Zhang, Changshui Zhang, *Member, IEEE*, Zhenwei Shi

***Abstract*— Source separation of musical signals is an appealing but difficult problem, especially in the single-channel case. In this paper, an unsupervised single-channel music source separation algorithm based on average harmonic structure modeling is proposed. Under the assumption of playing in narrow pitch ranges, different harmonic instrumental sources in a piece of music often have different but stable harmonic structures, thus sources can be characterized uniquely by harmonic structure models. Given the number of instrumental sources, the proposed algorithm learns these models directly from the mixed signal by clustering the harmonic structures extracted from different frames. The corresponding sources are then extracted from the mixed signal using the models. Experiments on several mixed signals, including synthesized instrumental sources, real instrumental sources and singing voices, show that this algorithm outperforms the general Nonnegative Matrix Factorization (NMF)-based source separation algorithm, and yields good subjective listening quality. As a side-effect, this algorithm estimates the pitches of the harmonic instrumental sources. The number of concurrent sounds in each frame is also computed, which is a difficult task for general Multi-pitch Estimation (MPE) algorithms.**

***Index Terms*— Single-channel Source Separation, Harmonic Structure, Multi-pitch Estimation, Clustering.**

I. Introduction

IN REAL music signals several sound sources such as a singing voice and instruments are mixed. The task of separating individual sources from a mixed signal is called sound source separation. This task interests researchers working on other applications such as information retrieval, automatic transcription and structured coding because having well-separated sources simplifies their problem domains.

Sound source separation problems can be classified by the number of sources and sensors. *Over-determined* and *determined* cases are those in which the number of sensors is larger than or equal to the number of sources, respectively. In these cases, Independent Component Analysis (ICA) [1]–[3] and some methods using source statistics [4], [8] can achieve good results. However, they encounter difficulties when handling *Under-determined* cases, in which sensors are fewer than sources. In these cases, some state-of-the-art methods employ source sparsity [5], [6] or auditory cues [7] to address the problem. The single-channel source separation problem is the extreme case of the under-determined source separation problem. Some methods which address this problem are reviewed in Section II.

According to the information used, sound source separation methods can be classified as *Supervised* and *Unsupervised*. Supervised methods usually need source solo excerpts to train individual source models [8]–[17], or overall separation model parameters [18], and then separate mixed signals using these models. Unsupervised methods [19]–[23], having less information to use, employ Computational Auditory Scene Analysis (CASA) [24], [25] cues, such as harmonicity, common onset and offset time, to tackle the separation problem. Also, nonnegativity [26], sparseness [4]–[6] and both [27] are employed by some unsupervised methods.

In this paper, we deal with the single-channel music source separation problem in an unsupervised fashion. Here each source is a monophonic signal, which has at most one sound at one time. It is found that in music signals, harmonic structure is an approximately invariant feature of a harmonic musical instrument in a narrow pitch range. Therefore, harmonic structures of these instruments are extracted from the spectrum of each frame of the mixed signal. We then learn Average Harmonic Structure (AHS) models, typical harmonic structures of individual instruments, by clustering the extracted structures, given the number of the instrumental sources. Using these models, corresponding sources are extracted from the mixed signal. We note that this separation algorithm needs not know the pitches of the sources. Instead, it gives Multi-pitch Estimation (MPE) results as a side-effect. The algorithm has been tested on several mixed signals of synthesized and real musical instruments as well as singing voices. The results are promising. The idea was first presented in [29]. This paper gives different formulations of estimating the F0s and extracting the harmonic structures, along with more detailed analysis, experiments and discussions.

The rest of this paper is organized as follows. Section II reviews some single-channel separation methods. The AHS model of music signals is proposed and analyzed in Section III. The model learning process and model-based separation process are described in Sections IV and V, respectively. Experimental results are illustrated in Section VI. We conclude with some discussions in Section VII.

This work is supported by the project (60475001) and (60605002) of the National Natural Science Foundation of China. Zhiyao Duan and Yungang Zhang contributed equally to this paper. The corresponding author is Zhiyao Duan (Email: duanzhiyao00@mails.tsinghua.edu.cn).

Zhiyao Duan and Changshui Zhang are with the State Key Laboratory on Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, China.

Yungang Zhang is with the Shanghai RS Technology CO., LTD, Shanghai, 200335, China.

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beijing University of Aeronautics and Astronautics, Beijing 100083, China.

II. Related work

The existing methods which aim at addressing the single-channel sound source separation problem can be classified into three broad and sometimes overlapping categories: Computational Auditory Scene Analysis (CASA)-based, spectral-decomposition-based and model-based methods [37].

A. CASA-Based Methods

CASA aims at using psychoacoustical cues [24], [25] to identify perceived auditory objects (e.g. partials of notes in music signals) and group them into auditory streams. Basic methods [19], [20], [23] use cues such as harmonicity, common onset and offset time and correlated modulation to characterize objects and build streams based on pitch proximity using binary masking [36]. Therefore, these methods can hardly separate sources playing the same pitch or having many overlapping partials.

To address this problem, a time-frequency smoothness constraint is added on the partials in [21], while spectral filtering techniques are used to allocate energy for overlapping partials in [22]. However, they both require knowledge of the pitches of the sources. In [30] some supervised information, such as timbre features learned on solo excerpts, are used to improve instrument separation.

B. Spectral-Decomposition-Based Methods

Similar to the "segmentation-grouping" process in CASA-based methods, spectral-decomposition-based methods first decompose the power or amplitude spectrogram into basis spectra vectors in a statistical fashion. These basis vectors are then clustered into disjoint sets corresponding to the different sources. Independent Subspace Analysis (ISA), which is an extension of ICA, is applied to the single-channel source separation problem [31]–[33]. Nonnegative Matrix Factorization (NMF) [34], constraining the basis vectors and/or time varying gains to be non-negative, has been found to efficiently decompose the spectrogram [12], [26], [27]. The sparseness constraint, which maintains consistency with the characteristics of note activities in music, is also added to the basis vectors and/or time varying gains in [5], [6], [28].

However, these methods generally encounter difficulties in the basis vectors clustering step. In [31] basis vectors are grouped by the similarity of marginal distributions, while in [32] instrument specific features are employed to facilitate the separation of drums. In [13] these features are learned from solo excerpts using Support Vector Machines (SVMs), but most other algorithms rely on manual clustering. In addition, these methods perform well on percussive instrument separation, but are rarely used with harmonic instruments and singing voices. In [12] vocals are separated from the accompanying guitar, but the vocal features are learned from solo excerpts.

C. Model-Based Methods

These methods usually establish generative models of the source signals to facilitate the separation task. In [9], [10], Hidden Markov Models (HMM) are trained on solo data and are factorially combined to separate the sources. In [15] a three-layer generative model is employed for Bayesian estimation of the sources. In [16], Bayesian harmonic models and perceptually motivated residual priors are employed, but this method concentrates primarily on decomposing signals into harmonic components without grouping them into source streams. In [17] a harmonic structure library is learned for each pitch of each instrument from individual note samples, and is then used to restore the overlapping partials in the separation step. However, this method requires that the pitches of the mixed signals be known.

These methods perform well on specific instrument separation problems, but have many model parameters to learn from solo excerpts. In addition, different recordings of the same instrument might change model parameters if the recording environment is changed. Therefore, such a prior assignment is not feasible. In [35] a spectral basis, which represents harmonic structure models of sources, is learned in an unsupervised fashion and is then used as NMF basis vectors to separate the signals. However, these bases are learned from the solo excerpts of the mixed signals, and fail when there is no solo data for each specific instrument, as described in [35].

D. Our Method

Our method is in essence a model-based method, which employs Average Harmonic Structure (AHS) models to separate harmonic instrumental sources from mixed signals. By borrowing ideas from CASA and spectral-decomposition-based methods, our method can deal with the problems encountered by each category of methods mentioned above. First, the AHS model is defined according to the harmonicity cues in CASA and represents the approximate invariant feature of a harmonic instrument. Second, the AHS models are learned directly from the mixed signals in an unsupervised way, so it does not need solo excerpts as training data. Third, when separating the signals, it manipulates the spectrum of each frame like the spectral-decomposition-based methods, but instead groups the components according to the AHS models. Therefore, it does not have difficulties grouping spectral components. Fourth, it allocates energy of overlapping partials based on the AHS models instead of binary masking, so that overlapping partial problems caused by sources in a harmonic relationship or the same pitch can be addressed.

III. Average Harmonic Structure Modeling for Music Signals

In the mixed signal, different sources usually have different timbre. Our motivation is to model the timbre characteristics to discriminate and separate the sources.

First consider the generation of sound from a harmonic source (such as a violin or a singer). Essentially, the sound is generated from a vibrating system (the violin string or the vocal cords) and then filtered by a resonating system (the violin body or the vocal tract) [39]. Although there is some coupling between the two systems [40], the source-filter model has been widely used in speech coding and music sound synthesis [41]. In the frequency domain, this process

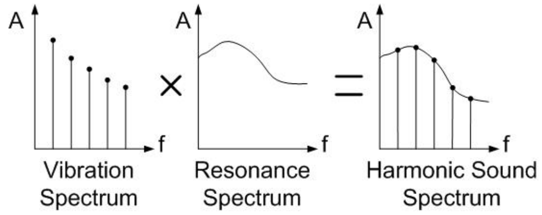


Fig. 1. The illustration of the generation process of a harmonic sound. The horizontal axis and the vertical axis are frequency and log-amplitude, respectively. The vibration spectrum is usually modeled as a series of harmonics with 6 or 12 dB/octave decrease in log-amplitude, while the resonance spectrum is modeled as a smooth curve representing the formants.

is illustrated in Fig. 1, where the spectrum of the harmonic source sound is the multiplication (addition in log-amplitude scale) of the spectrums of the two systems.

For an instrument, its nearly invariant feature when playing different pitches is its resonance spectrum, which can be modeled by its Mel-frequency Cepstral Coefficients (MFCC) [42]. This explains why MFCCs are so successful in instrument recognition for individual notes [38]. However, this feature is not suitable for source separation, because the MFCCs of each of the sources cannot be obtained from the mixed signals [43]. Therefore, a new feature that can characterize different sources and be easily obtained from the mixed signal is needed. The Average Harmonic Structure (AHS) is found a good choice.

Suppose $s(t)$ is a source signal (monophonic), which can be represented by a sinusoidal model [44]:

$$s(t) = \sum_{r=1}^{R} A_r(t) \cos[\theta_r(t)] + e(t) \tag{1}$$

where $e(t)$ is the noise component; $A_r(t)$ and $\theta_r(t) = \int_0^t 2\pi r f_0(\tau) d\tau$ are the instantaneous amplitude and phase of the r^{th} harmonic, respectively; $f_0(\tau)$ is the fundamental frequency at time τ; R is the maximal harmonic number. Although R is different for different sounds, it is set to 20 through this paper, since partials upper than 20 usually have very small amplitudes and are submerged in the sidelobes of the stronger partials, and for the notes having less than 20 partials, their upper partials are given a zero amplitude value.

Suppose that $A_r(t)$ is invariant in a short time (e.g. a frame), and is denoted as A_r^l in the lth frame; the harmonic structure in this frame is defined as the vector of dB scale amplitudes of the significant harmonics:

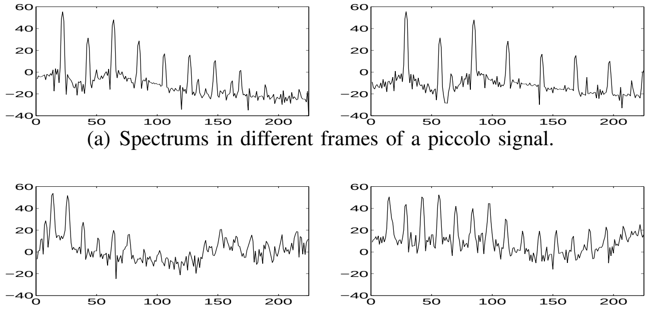
- Harmonic Structure Coefficient:

$$B_r^l = \begin{cases} 20\log_{10}(A_r^l), & \text{if} A_r^l > 1 \\ 0, & \text{otherwise} \end{cases}, r = 1, \ldots, R. \tag{2}$$

- Harmonic Structure:

$$\mathbf{B}^l = [B_1^l, \ldots, B_R^l] \tag{3}$$

The Average Harmonic Structure (AHS) model, just as its name implies, is the average value of the harmonic structures in different frames. Harmonic Structure Instability (HSI) is defined as the average variance of the harmonic structure coefficients.



(a) Spectrums in different frames of a piccolo signal.

(b) Spectrums in different frames of a voice signal.

Fig. 2. Spectrums of different signals. The horizontal axis is the number of frequency bins. The vertical axis is the log-amplitude in dB. Note that the difference in the log scale represents the ratio in the linear scale. The differences among the harmonics between the two spectrums of the piccolo signals are similar, while those of the vocal signals vary greatly. This shows that the piccolo signal has a stable harmonic structure while the vocal signal does not.

- Average Harmonic Structure (**AHS**):

$$\bar{\mathbf{B}} = [\bar{B}_1, \ldots, \bar{B}_R] \tag{4}$$

$$\bar{B}_i = \frac{1}{L_i} \sum_{l=1, B_i^l \neq 0}^{L_i} B_i^l, i = 1, \cdots, R. \tag{5}$$

- Harmonic Structure Instability (**HSI**):

$$HSI = \frac{1}{R} \sum_{i=1}^{R} \{ \frac{1}{L_i} \sum_{l=1, B_i^l \neq 0}^{L_i} (B_i^l - \bar{B}_i)^2 \} \tag{6}$$

where L_i is the total amount of frames where the ith harmonic structure coefficient is not 0.

Specifically, we use the AHS model for the following reasons: firstly, the AHS models are different for different sources, since the harmonic structures are determined from resonance characteristics, which are different for different sources. Secondly, harmonic structure is an approximately invariant feature for a harmonic instrument when it is played in a narrow pitch range [35]. Although the harmonics move up and down under a fixed profile, which is the resonance characteristic, their amplitudes will not change much if the pitch range is narrow enough, see Fig 2(a). Therefore, the average value AHS can be used to model the source. Thirdly, the harmonic structure of a singing voice is not as stable as that of an instrument, see Fig. 2(b). This is because the resonance characteristics for vocals vary significantly when different words are sung, causing the shape of the resonator (including the oral cavity) to change. This observation can be used to discriminate instrumental sounds from vocal sounds.

In calculating the AHS model, the harmonics $[A_1^l, \ldots, A_R^l]$ are obtained by detecting the peaks in the Short Time Fourier Transform (STFT) magnitude spectrum, as will be described in Section IV-A. The total power of all these harmonics is normalized to C, which can be an arbitrary constant. C is set to 100dB in this paper. Harmonic structure is defined in the log scale, simply because the human ear has a rough logarithmic sensitivity to signal intensity. Also in the log scale, the differences of the coefficients among the harmonics represent the power ratios in the linear scale, thus the Euclidean distance between two structures is meaningful. Note that the harmonic structure coefficient B_r^l is set to 0, if no significant corresponding harmonic is detected. The AHS is calculated in each dimension separately and only uses the non-zero coefficients. If there are too few (less than 30%) non-zero coefficients of a harmonic, in Eq. (4) the corresponding AHS value in that dimension is set to zero.

In Fig. 3, we calculated the AHS and HSI in the middle octave for several instruments in the Iowa musical instrument database [45]. We also calculated the AHS and HSI for four singing voice signals, where two are Italian voices downloaded from the Sound Quality Assessment Material (SQAM) website [46], and the other two are recorded Chinese voices.

From Fig. 3, it can be seen that the AHS of different instruments are different, although they are more similar for instruments in the same category (wood, brass or string). In addition, the HSI of instruments (especially brass instruments) are smaller than those of voices, even though the pitch range of the two female voices are narrower. Furthermore, for each instrument, in most cases the variances of different harmonics differ little. Therefore, we use the HSI to represent the variance of all the harmonics.

IV. AHS Model Learning from the Mixed Signal

For each source, an AHS model is learned directly from the mixed signals. The model learning algorithm consists of three steps: peak detection, harmonic structure extraction and harmonic structures clustering.

A. Peak Detection

In each frame, harmonics of sources are usually represented as peaks in the STFT spectrum, therefore, a peak detection step is essential. There are several peak detection algorithms in the literature, such as the cross-correlation method [47], which assumes that each peak has the shape of the spectrum of a sinusoid. It calculates the cross-correlation between the detected spectrum and the spectrum of a sinusoid, to find the peaks whose correlation values exceed a certain threshold. However, this method is not suitable in polyphonic music because many peaks do not resemble the spectrum of a sinusoid due to overlapping partials.

In a spectrum (the thin curve in Fig. 4), peaks are local maxima. However, it can be seen that there are many local maxima caused by side lobes or noise. We define significant peaks as those of interest relating to potential harmonics. We developed a detection method for finding these peaks. First, the smoothed log-amplitude envelope (the bold curve in Fig. 4) is calculated by convolving the spectrum with a moving Gaussian filter. Then the spectrum local maxima, which are higher than the envelope for a given threshold (e.g. 8 dB) are detected as significant peaks. Also, similar to [47], the peaks should be higher than a bottom line (the horizontal line in Fig. 4), which is defined as the maximum of the spectrum minus 50 dB. The bottom line can be seen as the noise floor, and the peaks under this line have negligible energy and high probabilities of being generated by noise or side lobes. Finally, the peak amplitudes and positions are refined by quadratic interpolation [48]. The detected peaks are marked by circles in Fig. 4.

The algorithm to detect significant peaks seems somewhat ad hoc, however, it provides robust peak detection results for the rest of the whole separation algorithm. The parameters of this algorithm used throughout this paper are the moving Gaussian filter, the 8-dB and the 50-dB thresholds, and we have found that our algorithm is not particularly sensitive to parameter settings. In fact, they can be replaced by other values, such as a moving average filter, 10-dB and 60-dB thresholds, without any change in separation performance.

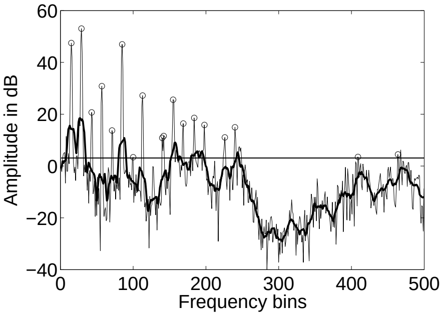


Fig. 4. Peak detection algorithm illustration. The thin curve is the spectrum, the bold curve is the smoothed log-amplitude envelope, and the horizontal line is the bottom line of the peaks. Detected peaks are marked by circles.

B. Harmonic Structure Extraction

Harmonic structures of each frame in the mixed signal are extracted from the peaks detected above. This process consists of two sub-steps. First, the number of concurrent sounds and the fundamental frequencies (F0s) are estimated. Second, the corresponding harmonics of F0s are extracted.

1) Maximum-Likelihood-Based F0s Estimation: For a particular frame of the mixed signal, suppose K peaks have been detected. Their frequencies and amplitudes are denoted as $f_1, f_2, \cdots, f_K$ and $A_1, A_2, \cdots, A_K$, respectively. Note that there can be multiple F0s, which we estimate using the Maximum Likelihood (ML) estimation with the spectral peaks as our observations.

Although the number of harmonic sources are known for the whole mixed signal, the number of concurrent sounds are unknown in each frame. Therefore, we estimate the F0s as well as the polyphony in each frame.

Suppose the polyphony in this frame is N, and the F0s are $f_0^1, f_0^2, \cdots, f_0^N$. The likelihood function can be formulated as:

$$\begin{aligned} p(O|f_0^1, f_0^2, \cdots, f_0^N) &= p(f_1, f_2, \cdots, f_K|f_0^1, f_0^2, \cdots, f_0^N) \\ &= \prod_{i=1}^{K} p(f_i|f_0^1, f_0^2, \cdots, f_0^N) \end{aligned} \quad (7)$$

where O is the observation, and is represented by the frequencies of the peaks, because it contains the most information that can be used at this point. It is also assumed that the frequencies of the peaks are conditionally independent given the F0s. This is reasonable and is commonly treated in the spectral probabilistic modeling literature [49].

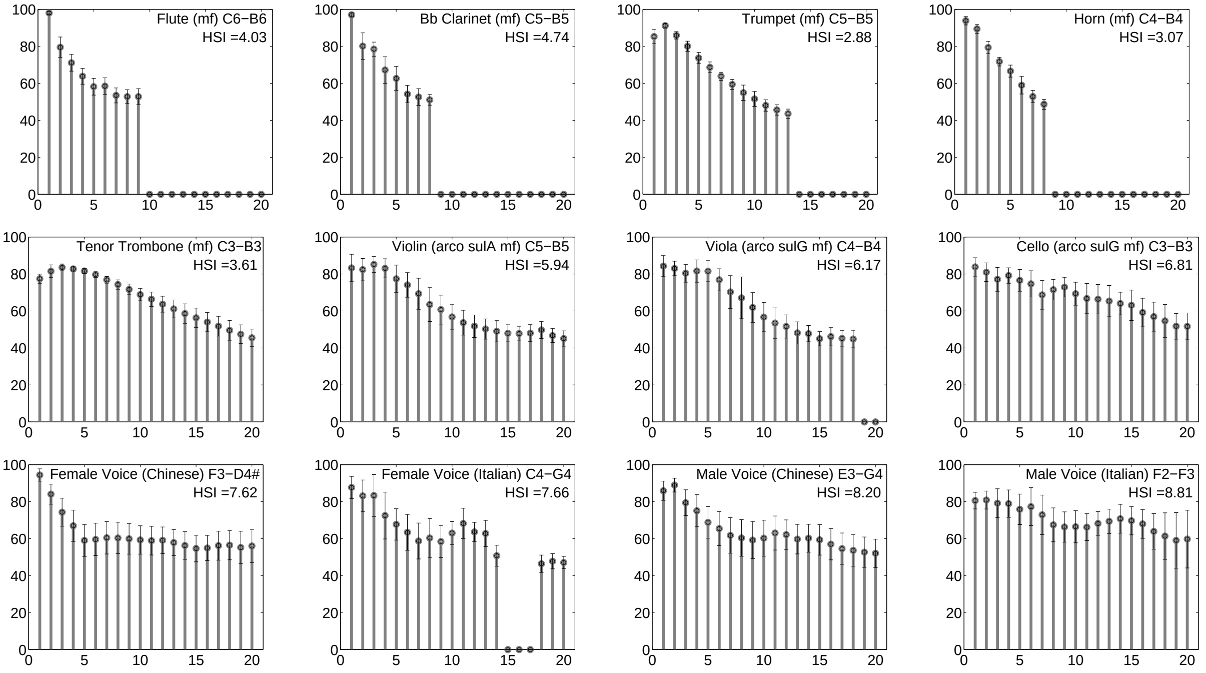


Fig. 3. The AHS models of several harmonic instrumental and vocal signals in specific dynamic ranges and pitch ranges. The horizontal axis and the vertical axis are the harmonic number and the log-amplitude in dB, respectively. Zero in the AHS value means that there is no significant corresponding harmonic detected. The twice standard deviation of each harmonic is depicted as the small vertical bar around the AHS value.

For modeling of the likelihood of a peak f_i, $d(f_i, f_0^j)$ is calculated, which is the frequency deviation between peak f_i and the corresponding ideal harmonic of fundamental f_0^j. The likelihood is modeled as a Gaussian distribution of $d(f_i)$, which is defined as the smallest frequency deviation $d(f_i, f_0^j)$ among all the F0s, to follow the assumption that each peak is generated by the nearest F0.

$$p(f_i|f_0^1, f_0^2, \cdots, f_0^N) = \frac{1}{C_1}\exp\{-\frac{d^2(f_i)}{2\sigma_1^2}\} \tag{8}$$

$$d^2(f_i) = \min_j d^2(f_i, f_0^j) \tag{9}$$

$$d(f_i, f_0^j) = \frac{f_i/f_0^j - [f_i/f_0^j]}{[f_i/f_0^j]} \tag{10}$$

where $[\cdot]$ denotes rounding to the nearest integer. σ_1 is the standard deviation and is set to 0.03 to represent half of the semitone range. C_1 is the normalization factor.

Note that if a new fundamental frequency f_0^{N+1} is added to the existing F0s, the likelihood function will increase because of the minimum operation. Thus, the likelihood of Eq. (7) approaches $\frac{1}{C_1^K}$, as the number of F0s goes towards infinity.

$$p(O|f_0^1, \cdots, f_0^N) \leq p(O|f_0^1, \cdots, f_0^N, f_0^{N+1}) \tag{11}$$

This is the typical overfitting problem of ML method and can be addressed by applying model selection criterions. Here the Bayesian Information Criterion (BIC) [50] is adopted to estimate the number of concurrent sounds N.

$$BIC = \ln p(O|f_0^1, f_0^2, \cdots, f_0^N) - \frac{1}{2}N\ln K \tag{12}$$

TABLE I

Algorithm flow of the multiple F0s estimation

1) Set $N = 1$, calculate and store f_0^1 which maximize Eq. (7);
2) $N = N + 1$;
3) Calculate and store f_0^N, which maximize $p(O|f_0^1, \cdots, f_0^{N-1}, f_0^N)$;
4) Repeat 2-3 until $N = 10$;
5) Select a value for N which maximize Eq. (12). The estimated F0s are $f_0^1, \cdots, f_0^N$.

The number of F0s and their frequencies are searched to maximize Eq. (12). In order to reduce the search space and to eliminate the trivial solution that the estimated F0s are near 0, the F0s are searched around the first several peaks. This also eliminates some half-fundamental errors. However, the search space is still a combinatorial explosion problem. Hence, we use a greedy search strategy, which starts with $N = 1$. The algorithm flow is illustrated in Table I.

In this algorithm, the number of concurrent sounds N and F0s may not be estimated exactly, but the results are satisfactory for the harmonic structure clustering step, and the final, correct F0s will be re-estimated in Section V-A.

It is noticed that the idea of probabilistic modeling of the STFT peaks has been proposed by Thornburg and Leistikow *et al.*. In [51] they aims at melody extraction and onset detection of the monophonic music signal, and in [52] they aims at chord recognition from a predefined codebook consisting 44 kinds of chords of the polyphonic music signal. However, both of them do not handle the general multiple F0 estimation problem.

2) Harmonics Extraction: After the F0s have been estimated, the corresponding harmonics are extracted from the nearest peaks in the mixture spectrum, with the constraint that the deviation in Eq. (9) lies in $[-0.03, 0.03]$. The log-amplitudes of the peaks are used to form the harmonic structure in Eq. (3). If there is no peak satisfying this constraint, the harmonic is assumed missing and the log-amplitude in Eq. (3) is set to 0. Also the number of non-zero values in a harmonic structure should be more than 5, based on the observation that a harmonic instrumental sound usually have more than 5 harmonics. This threshold is invariant for all the experiments.

Note that in the spectrum of a polyphonic signal, the harmonics of different notes often coincide. The amplitudes of some peaks are influenced collectively by the coincident harmonics. Therefore, the extracted harmonic structure is not exact. However, because the relationship of the notes varies in different frames, the way in which harmonics coincide also varies. For example, suppose the r^{th} harmonic coincide in one frame, but in the other frames it may not coincide. We can still learn the amplitude of the r^{th} harmonic from all the frames. This is the motivation behind the harmonic structures clustering algorithm described in Section IV-C.

Special case: In the case that one source is always one octave or several octaves higher than another source, as in Section VI-B, the multiple F0 estimation and the harmonic extraction method above cannot detect the higher F0 and extract its harmonics. This is because of the following reasons. First, in each frame, though the harmonics of the octave(s)-higher F0 cannot be entirely overlapped by those of the lower F0 due to slight detuning, and the likelihood will increase after adding the octave(s)-higher F0, the increase is little and not enough to compensate the decrease in Eq. (12) caused by the model complexity penalty, since the detuning is much smaller than σ_1 in the Gaussian model in Eq. (8). One might consider to change the weight of the model complexity penalty, but it is difficult, because the design of the penalty should also consider eliminating the false F0s which may be caused by false peaks due to noise and side lobes. Therefore, the balance is hard to achieve and the octave(s)-higher F0 cannot be detected. Second, because this octave(s) relationship happens in all the frames, the F0s and their harmonics of the octave(s)-higher source are always not detected. Note that if this octave(s) relationship just happens in some but not all frames, the harmonics of the higher source can still be detected somewhat and used to learn its AHS model. We address this special case by separately estimating the M most likely F0s, using Eq. (7) with $N = 1$, where M is the number of sources and given as prior. This method emphasizes the F0 candidates as long as at whose harmonic positions some peaks occur. Therefore, some harmonics of the true F0s will be also detected as false F0s. In order to eliminate these errors, a constraint is adopted that an extracted harmonic structure should not be a substructure of any others. This constraint discards the false F0s mentioned above, because the harmonic structures of these false F0s are always some substructures of the true F0s; but it does not prevent the detection of the true, octave(s)-higher F0, because its harmonics are not exactly overlapped by those of the octave(s)-lower F0 due to the slight detuning, and the harmonic structures of the higher F0 are not substructures of the lower F0.

C. Harmonic Structures Clustering

After the harmonic structures of all F0s in all frames have been extracted, a data set is obtained. Each data point, a harmonic structure, is a 20-dimensional vector. The distance between two data points is measured in the Euclidean sense. As analyzed in Section III, harmonic structures are similar for the same instrument, but different for different instruments. Therefore, in this data set there are several high density clusters, which correspond to different instruments, respectively. In addition, these clusters have different densities, because the stabilities of the harmonic structures of different instruments are not the same. Furthermore, harmonic structures of a singing voice scatter like background noise, because they are not stable.

In order to learn the AHS model of each instrument, an appropriate clustering algorithm should be used to separate the clusters corresponding to the instruments, and remove background noise points corresponding to the singing voice. The NK algorithm [54] is a good choice for this application. Its basic idea is to calculate at each point the sum of the covariance of its neighborhood (K nearest neighbors), and use this value to represent the inverse of the local density at the point. Though there is no explicit equation between this value and the local density, the larger the value is, the lower the local density is. The point whose this value is larger than the average value of its neighbors for one standard deviation is assumed to be a background noise point, and is removed from the data set, since it is a relatively low density point. The remaining points connect to their neighbors and form disconnected clusters. Since this algorithm only focuses on relative local densities, it can handle the data set that consists of clusters with different shapes, densities, sizes and even some background noise. The number of neighbors K is the only adjustable parameter in this algorithm for this application. It decides how many disconnected clusters will be formed. The bigger K is, the fewer clusters are formed. In our experiments, the number of sources is used to guide the choice of K.

We note that AHS models of only the harmonic instrumental sources can be learned from the clustering process, while those of the inharmonic or noisy sources (such as a singing voice) cannot be learned.

V. Music Signal Separation Based on AHS Models

This section discusses how to separate the sources in the mixed signal by using the learned AHS models. The basic idea is to re-estimate the F0 corresponding to the AHS model in each frame using ML estimation, then re-extract the harmonics from the mixture spectrum and reconstruct the time domain signal by using the Inverse Fast Fourier Transform (IFFT).

A. Maximum-Likelihood-Based F0 Re-estimation

Compared with the ML-based F0s estimation algorithm in Section IV-B, here, a single-F0 estimation algorithm, given an AHS model is used. The likelihood is formulated as follows:

$$p(O|f,\bar{\mathbf{B}}) = p(f_1,\cdots,f_K,A_1,\cdots,A_K|f_0,\bar{\mathbf{B}}) = \prod_{i=1}^{K} p(f_i,A_i|f_0,\bar{\mathbf{B}}) \quad (13)$$

where O denotes the observation (the spectrum of a frame), $f_1,\cdots,f_K$ and $A_1,\cdots,A_K$ are the frequencies and amplitudes of the peaks, respectively. $\bar{\mathbf{B}}$ is the AHS model and f_0 is its corresponding fundamental frequency.

Compared with Eq. (7), additional information about the amplitudes of the peaks is added to represent the observation, since amplitude information is contained in the AHS model. The frequencies and amplitudes of the peaks are still assumed independent given the F0 and the AHS model as before.

The likelihood of each peak is derived using the chain rule and the independence between the frequency of the peak and the AHS model.

$$\begin{aligned} & p(f_i,A_i|f_0,\bar{\mathbf{B}}) \\ = \;& p(f_i|f_0,\bar{\mathbf{B}}) \cdot p(A_i|f_i,f_0,\bar{\mathbf{B}}) \\ = \;& p(f_i|f_0) p(A_i|f_i,f_0,\bar{\mathbf{B}}) \\ = \;& \frac{1}{C_2} \exp\{-\frac{d^2(f_i,f_0)}{\sigma_1^2}\} \exp\{-\frac{D^2(A_i,\bar{\mathbf{B}})}{\sigma_2^2}\} \end{aligned} \quad (14)$$

where $p(f_i|f_0)$ is modeled as a Gaussian distribution of $d(f_i,f_0)$, which is the frequency deviation of f_i from the nearest ideal harmonic of f_0 as before. σ_1 is the standard deviation and is still set to 0.03 typically, to represent the half of the semitone range. $p(A_i|f_i,f_0,\bar{\mathbf{B}})$ is modeled as a Gaussian distribution of $D(A_i,\bar{\mathbf{B}})$, which is the log-amplitude deviation of A_i from the nearest ideal harmonic $\bar{B}_{[f_i/f_0]}$). σ_2 is set to the HSI of the AHS model. $[\cdot]$ denotes rounding to the nearest integer. C_2 is the normalization factor.

$$d^2(f_i,f_0) = \min\left(\left(\frac{f_i/f_0 - [f_i/f_0]}{[f_i/f_0]}\right)^2, 4\sigma_1^2\right) \quad (15)$$

$$D^2(A_i,\bar{\mathbf{B}}) = \min\left((A_i - \bar{B}_{[f_i/f_0]})^2, 4\sigma_2^2\right) \quad (16)$$

Note that the minimum operation in these two equations represent that, if the peak f_i lies outside the semitone range of the ideal harmonic of f_0, or the log-amplitude of the peak deviates more than twice the standard deviation, it is assumed that the peak is not generated by f_0. Therefore, the frequency and the log-amplitude deviations of this peak from the ideal harmonic of the F0 should be limited to avoid over-penalizing in the likelihood function.

After all the F0s corresponding to the AHS model in all the frames have been estimated, two cascade median filters with length 3 and 7 are employed to the F0 line to eliminate abrupt errors.

B. Re-extraction of Harmonics

For each estimated F0, the corresponding observed harmonics are extracted from the mixture spectrum to form the harmonics of the reconstructed source spectrum. If the normalized log-amplitude (see Section III) of a harmonic in the mixture spectrum deviates less than σ_2 in Eq. (14), it is used in the separated source spectrum; otherwise the value in the AHS model is used. Note that for reconstructing the harmonic sources, this process can either be a cascade one that the harmonics are extracted one by one and subtracted from the mixture before estimating further sources, or a parallel one that all the harmonic sources are extracted directly from the mixture spectrum. However, for reconstructing the inharmonic or noisy sources, a cascade process is required that the harmonics of all the harmonic sources are removed, with the residual spectrum being left.

The extraction results eliminate many errors caused in the first extraction described in Section IV-B. The reason is that in the former extraction step, only information from one frame is used. However, in the re-extraction step, the AHS models are used, which contain information of all frames. Fig. 5 illustrates the improvements of harmonics extraction made by using the AHS model. Fig. 5(a) and (b) are the spectrums of two instrument sources in one frame. Fig. 5(c) and (d) are the preliminary harmonic structure extraction results (marked by circles) from the mixed signal, corresponding to estimated F0s of the two sources. In Fig. 5(c), the last extracted harmonic actually belongs to the second source but is assigned to the first source. Also, in Fig. 5(d), the 3rd, 5th, 7th, 8th and 9th extracted harmonics belong to the first source but are assigned to the second source. These errors are caused by the incorrect F0s being estimated using only the frequency likelihood of the peaks in Eq. (7). In contrast, Eq. (13), the re-estimation of F0s using the AHS models (Fig. (5(e) and (f)), incorporates additional information about the log-amplitude likelihood of the peaks, which eliminates all of the harmonic extraction errors, see Fig. 5(g) and (h).

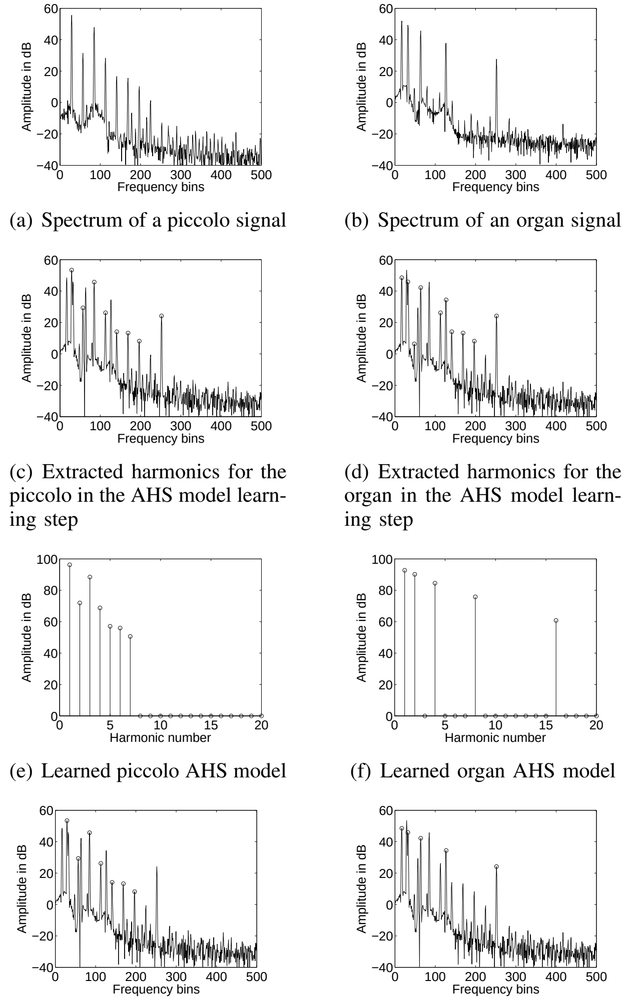
In addition, often some harmonics of different sources overlap and their amplitudes are difficult to estimate. In this case, the AHS model helps determine their amplitudes, without using other spectral filtering techniques [22].

C. Reconstruction of the Source Signal

In each frame, for each instrumental source, the harmonics extracted from the mixed signal form a new magnitude spectrum. To get the complex spectrum, the phases of the mixed signal or those estimated from a phase generation method [55] can be used. The waveform of each source is reconstructed by performing the inverse FFT of the complex spectrum and using an overlap-add technique. The waveform of the inharmonic or noisy source (such as a singing voice or a drum) is synthesized from the residual spectrum after extracting the harmonics. In each frame, the energies of the sources are all normalized to that of the mixed signal.

In most cases, the use of the mixed signal phases produces good results. However, if the original phases are not suitable for the separated sources, the resulting waveform may become distorted because of discontinuities at frame boundaries. These distortions are attenuated by the overlap-add procedure.

Note that our algorithm can deal with several harmonic sources mixed with only one inharmonic or noisy source,



(a) Spectrum of a piccolo signal

(b) Spectrum of an organ signal

(c) Extracted harmonics for the piccolo in the AHS model learning step

(d) Extracted harmonics for the organ in the AHS model learning step

(e) Learned piccolo AHS model

(f) Learned organ AHS model

(g) Re-extracted harmonics using the piccolo AHS model

(h) Re-extracted harmonics using the organ AHS model

Fig. 5. Harmonics extraction and re-extraction results. The harmonics extraction accuracy is significantly improved by using the AHS models.

because the harmonic sources are extracted from the mixed signal using AHS models, leaving the inharmonic or noisy source in the residual.

VI. Experimental Results

The proposed algorithm has been tested on several mixed signals, including synthesized and real harmonic instrument signals, and singing voices. All these sounds have a sampling frequency of 22050 Hz, and are analyzed using 93 ms Hamming window with a 46ms hop-size, to get constant overlap reconstruction (COLA) [48]. The experimental results including audio files are accessible at http://mperesult.googlepages.com/musicseparationresults.

The performance of the experiments are all measured using the evaluation method proposed in [56]. This gives the overall signal to distortion ratio (SDR), the signal to interference ratio (SIR), i.e. the ratio of the true source to the interference of the other sources, and the signal to artifact ratio (SAR) i.e. a measure of the artifacts introduced by the method. Essentially, the estimated source is decomposed into a true source part plus error terms corresponding to interferences, additive noise and algorithmic artifacts, by projecting the estimated source to the corresponding signal space. The energy ratios of these terms form the definitions of SDR, SIR and SAR. These values are calculated using BSS_EVAL toolbox [57].

TABLE II
Performance measurement of separating two synthesized harmonic instrumental sources.

	Piccolo			Organ		
	AHS	NMF	Oracle	AHS	NMF	Oracle
SDR	14.2	11.3	15.9	11.8	9.0	14.1
SIR	27.9	20.1	28.7	25.1	20.6	24.9
SAR	14.4	11.9	16.1	12.1	9.3	14.5

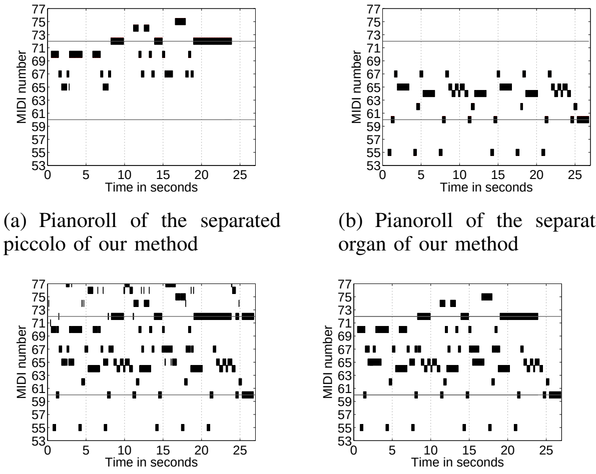
For comparison, the oracle separation results are calculated using BSS_Oracle toolbox [58]. Note that the oracle results are the theoretically, highest achievable results of the time-frequency masking-based methods, e.g. [9], [14], [15], which are usual methods used for single-channel source separation problems. The oracle results can only be obtained when the reference sources are available. Therefore, it serves as an upper bound on the performance. In addition, we implemented the NMF-based source separation algorithm [26]. In this algorithm, the spectrogram of each source signal is decomposed into 15 components, which span a subspace. The spectrogram of the mixed signal is decomposed into 30 components. The components of the mixed signal are clustered by distance measurements in the subspaces corresponding to the sources, and classified into the closest subspace. Finally, the components in the same subspace are used to synthesize the corresponding source. Note that this NMF-based method requires reference sources.

A. Two Synthesized Harmonic Instruments

The two MIDI-synthesized sources are played using piccolo and organ patches, respectively, and are mixed by addition with approximately equal energy without noise. The learned AHS models are illustrated in Fig. 5(e) and (f). Since the mixture is noise free and the two sources both have an AHS model, we have three methods to separate the two sources: 1) Extracting the piccolo source using its AHS model and leaving the organ source in the residual; 2) Extracting the organ source and leaving the piccolo source; 3) Extracting the two sources both from the mixed signal using their own AHS models. The performances of the three methods are similar, and the first method is used in this section.

The numerical comparisons of the results are listed in Table II. It can be seen that the SDRs and SARs of our algorithm still have some room to improve to the oracle results, while the SIRs of our algorithm approach or even outperform those of the oracle results. This is probably because our algorithm is not a binary masking method, which allocates the energy of a time-frequency point to only one source. Our algorithm allocates the overlapping partials to both sources according to their own AHS models. In addition, our algorithm outperforms the NMF-based method on all the indices. This is promising, since this NMF-based method uses the reference source signals to guide the separation, while our method does not.

Our method also provides the Multi-pitch Estimation (MPE) results as side-effects. The MPE pianorolls are illustrated in Fig. 6(a) and (b). For comparison, we calculated the MPE results in Fig. 6(c) using the current state-of-the-art MPE algorithm [53], which estimates the pitches of each frame in an iterative spectral subtraction fashion and has been found to be successful in individual chord MPE tasks. The polyphony number is estimated using the recommended method in that paper with the restriction that the number not exceed 2 as a prior, so that our algorithm and [53] are given the same information. The true pianoroll is illustrated in Fig. 6(d). All the pianorolls are painted using the MIDI Toolbox [59].



(a) Pianoroll of the separated piccolo of our method

(b) Pianoroll of the separated organ of our method

(c) Pianoroll of the MPE [53] results

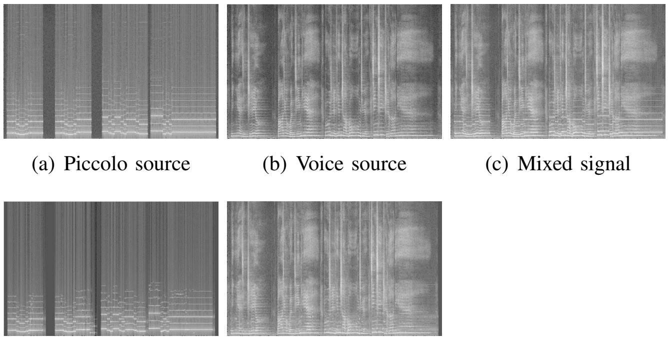
(d) Pianoroll of the true MPE results

Fig. 6. Comparison of the MPE results between [53] and our method

In Fig. 6, it can be seen that our algorithm gives good MPE results for both sources, except for several occasional errors at note transitions. Furthermore, on this specific mixture, our algorithm outperforms [53] in several regards. First, it correctly determines which note belongs to which source, while this task cannot be accomplished by MPE algorithms. Second, it gives better estimations of pitch numbers in each frame compared to [53]. For example, in the intermission of one source (such as the 5th-6th second, and the 24th-27th second of the piccolo source), there is actually only one pitch in the mixed signal. Our algorithm correctly estimates the only pitch at that moment, while [53] incorrectly adds a note. Third, it deals well with the overlapping note cases. For example, the short note of MIDI number 65 at about the 2nd-3rd second of the piccolo source is entirely overlapped by the long note of the organ source. Our algorithm correctly estimates the two notes, while [53] adds a false note at MIDI number 77.

B. A Synthesized Harmonic Instrument and A Singing Voice

The instrumental source is the piccolo signal used in Section VI-A, and the singing voice is a Chinese female vocal which is one octave below. The mixed signal is generated by adding the two sources with equal energy and without noise.



(a) Piccolo source (b) Voice source (c) Mixed signal

(d) Separated piccolo (e) Separated voice

Fig. 7. Spectrograms of a synthesized harmonic instrumental source, a singing voice, the mixture and the separated signals. The x-axis is time from 0 to 27 seconds, the y-axis is linear frequency from 0 to 11025Hz. The intensity of the graph represents the log-amplitude of the time-frequency components, with white representing high amplitude.

TABLE III

PERFORMANCE MEASUREMENT OF SEPARATING A HARMONIC INSTRUMENTAL SOURCE AND A SINGING VOICE.

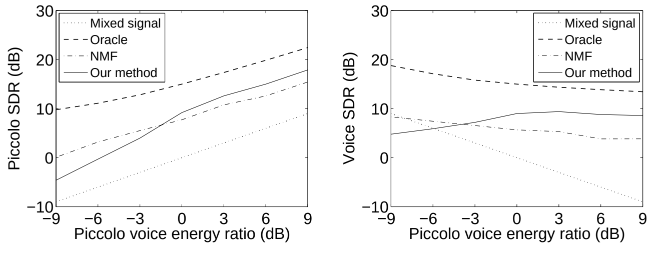
	Piccolo			Vocal		
	AHS	NMF	Oracle	AHS	NMF	Oracle
SDR	9.2	7.7	15.0	9.0	5.6	15.0
SIR	19.7	17.8	27.7	30.8	15.5	23.0
SAR	9.7	8.3	15.3	9.1	6.2	15.6

Note that the one octave relationship is the special case mentioned in Section IV-B, where the F0s are firstly estimated using the single F0 estimation algorithm, and the sub-structure elimination mechanism is employed to avoid F0 and harmonic structure errors. Fig. 7 illustrates the spectrograms of the sources, mixtures and the separated signals. It can be seen that the separated signals are similar to the sources, except that some higher harmonics of the piccolo signal are not preserved. This is because these harmonics are hard to detect in the mixed signal, and cannot be learned in the AHS model.

The SDR, SIR and SAR values are listed in Table III. It can be seen that our method outperforms the NMF method in all the indices. Compared with the oracle results, there is still some room for improvement, though, our SIR of the voice source is higher indicating that the components of the piccolo signal better extracted.

In order to inspect the performance comparison more deeply, we mixed the two sources with different energy ratios, and depicted the SDR curves of the mixed signal, oracle results, NMF results and our results, versus the energy ratio between the piccolo source and the voice source as in Fig. 8.

In Fig. 8(a), the SDR curve of the mixed signal represents the SDRs of the mixed signal viewed as the estimated piccolo source, therefore, its value equals to the energy ratio on the abscissa. Similarly, in Fig. 8(b) the SDR curve is inverse proportional to the abscissa. The two curves are the baselines, where nothing has been done to the mixed signal. The oracle lines are the highest ones. They are generally the theoretical upper bounds of the single-channel source separation performance. The piccolo source's oracle SDR increases with the



(a)

(b)

Fig. 8. SDRs vary with the energy ratio between the piccolo and the voice.

energy ratio between the piccolo and the voice, while the voice source's oracle SDR decreases. It indicates that the extraction of a source is easier when its source's energy is larger in the mixed signal. The middle two lines in each sub-figure are our results and the NMF results. It can be seen that when the piccolo voice energy ratio is lower than -3dB, the performance of our method is worse than that of the NMF method, because in this case the spectrums of the piccolo signal are submerged by those of the voice signal, thus the AHS model is hard to obtain and the piccolo signal is hard to extract using the AHS model. However, our method generally outperforms the NMF method, especially when the piccolo energy is large.

In our algorithm, the likelihood function in Eq. (13) is the key formula to re-estimate the F0s. Its maximum among all the possible F0s gives the likelihood of the observation given an AHS model. Here given the AHS model of the piccolo signal, the minus log-likelihood of the piccolo signal and the voice signal are calculated separately in Fig. 9.

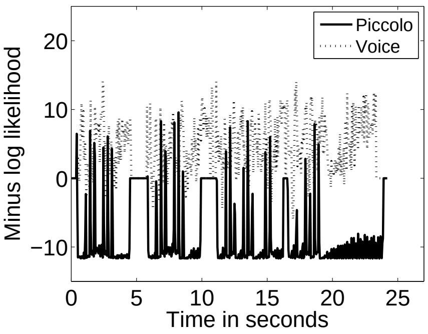


Fig. 9. The maximum of the minus log likelihood (Eq. (13)) among all the F0s, given the AHS model of the piccolo signal.

From Fig. 9, it can be seen that the minus log likelihood of the piccolo signal is much smaller than that of the singing voice. It means two things: First, the AHS model learned from the mixed signal correctly represents the characteristic of the piccolo source. Second, the likelihood definition of Eq. (13) is proper that it discriminates the piccolo source and the singing voice distinctly, and guarantees separation performance. In addition, it can be seen that the minus log likelihood of the piccolo signal varies with time. For most of the time, it is rather small, however, at note transitions (refer to Fig. 6(a)) it is large. This is because the harmonic structures are not stable and deviate from the AHS model somewhat at note transitions. This indicates that the AHS model better suits stationary phases versus transitory phases.

TABLE IV

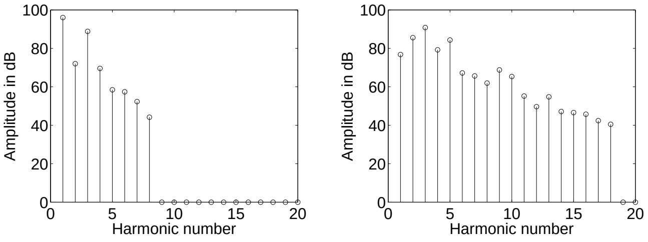
PERFORMANCE MEASUREMENT OF SEPARATING TWO SYNTHESIZED HARMONIC INSTRUMENTS AND A SINGING VOICE.

	Piccolo			Oboe			Voice		
	AHS	NMF	Oracle	AHS	NMF	Oracle	AHS	NMF	Oracle
SDR	11.2	13.7	19.4	10.1	12.9	17.3	7.6	6.2	16.9
SIR	23.1	22.5	34.4	33.1	26.6	31.1	23.1	28.6	30.9
SAR	11.5	14.4	19.5	10.2	13.1	17.5	7.7	6.2	17.1

C. Two Synthesized Harmonic Instruments and a Voice

The three sources and their pitch ranges are a piccolo (F4-D5♯), an oboe (G3♯-A4♯) and a male voice (G2-G3), respectively. Different from Section VI-B, the three sources are not related, and are mixed with the energy ratio of piccolo to oboe 2.5dB, piccolo to voice 6.7dB.

The two learned AHS models of the piccolo and the oboe are depicted in Fig. 10. As described in Section VI-B, the separation performance is better if the source with the biggest energy is extracted first. Therefore, here we first extract the piccolo signal using its AHS model, then extract the oboe signal from the residual. The final residual is the voice signal. The numerical results are listed in Table IV. From this table, it can be seen that the performance of the AHS method and the NMF method are similar, and both the two methods still have much room for improvement compared to the oracle results.



(a) Learned piccolo AHS model

(b) Learned oboe AHS model

Fig. 10. AHS models learned from the mixed signal.

D. Two Real Harmonic Instruments

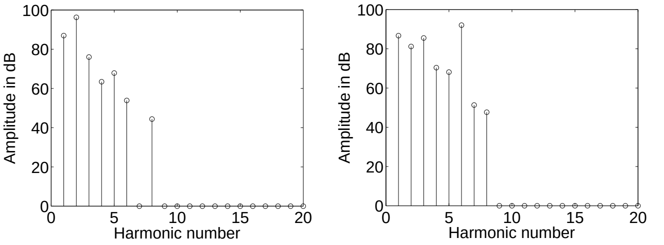
The two instrumental sources are oboe (E5-F5) and euphonium (G3-D4) solo excerpts, extracted from unrelated commercial CDs, however, they have a harmonic relationship. They also have some vibrato and reverberation effects. The mixed signal is generated by adding the two sources without noise, where the energy ratio is 2.3dB (Euphonium to Oboe).

The two corresponding AHS models learned from the mixed signal are depicted in Fig. 11. As described in Section VI-A, there are three methods to separate the two harmonic instrumental sources using the two AHS models. However, it is found that the results achieved by first extracting the oboe source and leaving the euphonium source as the residual are superior, though the energy of the Euphonium is larger. This

TABLE V
PERFORMANCE MEASUREMENT OF SEPARATING TWO REAL HARMONIC INSTRUMENTAL SOURCES.

	Oboe			Euphonium		
	AHS	NMF	Oracle	AHS	NMF	Oracle
SDR	8.7	7.9	25.8	4.6	2.3	18.9
SIR	25.8	10.2	41.1	14.5	9.0	35.4
SAR	8.8	12.0	26.0	5.3	3.8	19.0

is likely the case that the euphonium AHS model was not learned well. As shown in Fig. 11, the learned 6th harmonic is significantly higher than the other harmonics, which is not usual. The reason for this abnormality is because the pitches are lower, such that the harmonics of the euphonium are contaminated more severely by those of the oboe signal.



(a) Learned oboe AHS model (b) Learned euphonium AHS model

Fig. 11. AHS models learned from the mixed signal.

The numerical evaluation results are listed in Table V. It can be seen that there is still much room for improvement comparing our results to those of the oracle results. However, our results are better than those of the NMF-based method except the SAR value of the oboe signal. This is being the case because some artifacts are introduced by the AHS model when extracting the oboe signal. However, our SIR values are significantly higher, illustrating that the AHS model is better at suppressing interference.

In addition to these experiments, more are accessible at http://mperesult.googlepages.com/musicseparationresults.

VII. Conclusion and Discussion

In this paper, an unsupervised model-based music source separation algorithm is proposed. It is found that the harmonic structure is an approximately invariant feature of a harmonic instrument in a narrow pitch range. Given the number of instrumental sources and the assumption that the instruments play in narrow pitch ranges, the Average Harmonic Structure (AHS) models of different instruments are learned directly from the mixed signal, by clustering harmonic structures extracted from different frames. The AHS models are then used to extract their corresponding sources from the mixed signal. Experiment separating synthesized instrumental sources, real instrumental sources and singing voices, show that the proposed method outperforms the NMF-based method, which serves as a performance reference for the separation task.

The proposed algorithm also estimates the pitches of the instrumental sources as a side-effect. It can automatically decide the number of concurrent sounds and identify the overlapped notes, which is difficult for general Multi-pitch Estimation algorithms.

It is noticed that the proposed algorithm cannot handle a mixed signal which has more than one inharmonic or noisy sources (such as drums and singing voices), because these sources cannot be represented by the AHS model and are left in the residual during separation.

For future work, we would like to extend the AHS model to model properly stringed-instruments by adding the time dimension, since harmonic structures of these instruments vary with time. Finally, modeling the resonant features of instrumental sources can better characterize instruments and be more robust against pitch variations.

VIII. Acknowledgment

We thank A. Klapuri for providing his Multi-pitch Estimation program for comparison. We thank Nelson Lee and David Yeh for their improvement of the presentation. We also thank the editor and three reviewers for their very helpful comments.

References

[1] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural Networks*, vol. 13, pp. 411-430, 2000.

[2] L. C. Parra and C. V. Alvino, “Geometric source separation: merging convolutive source separation with geometric beamforming,” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 352-362, 2002.

[3] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21-34, 1998.

[4] M. Zibulevsky, P. Kisilev, Y. Y. Zeevi and B. Pearlmutter, “Blind source separation via multinode sparse representation,” in *Proc. NIPS*, 2002.

[5] Te-Won Lee, M. S. Lewicki, M. Girolami and T. J. Sejnowski, “Blind source separation of more sources than mixtures using overcomplete representations,” *IEEE Signal Process. Lett.*, vol. 6, no. 4, pp. 87-90, 1999.

[6] C. Fevotte and S. J. Godsill, “A Bayesian approach for blind separation of sparse sources,” *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 6, pp. 2174-2188, 2006.

[7] H. Viste and G. Evangelista, “A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures,” *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 3, pp. 1051-1061, 2006.

[8] M. J. Reyes-Gomez, B. Raj and D. Ellis, “Multi-channel source separation by factorial HMMs,” in *Proc. ICASSP*, 2003, pp. I-664-667.

[9] S. T. Roweis, “One microphone source separation,” in *Proc. NIPS*, 2001, pp. 15-19.

[10] J. Hershey and M. Casey, “Audio-visual sound separation via hidden markov models,” in *Proc. NIPS*, 2002.

[11] Gil-Jin Jang and Te-Won Lee, “A probabilistic approach to single channel blind signal separation,” in *Proc. NIPS*, 2003.

[12] S. Vembu and S. Baumann, “Separation of vocal from polyphonic audio recordings,” in *Proc. ISMIR*, 2005.

[13] M. Helén and T. Virtanen, “Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,” in *Proc. EUSIPCO*, 2005.

[14] L. Benaroya, F. Bimbot and R. Gribonval, “Audio source separation with a single sensor,” *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 1, 191-199, 2006.

[15] E. Vincent, “Musical source separation using time-frequency source priors,” *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 1, pp. 91-98, 2006.

[16] E. Vincent and M. D. Plumbley, “Single-channel mixture decomposition using Bayesian harmonic models,” in *Proc. ICA*, 2006, pp. 722-730.

[17] M. Bay and J. W. Beauchamp, “Harmonic source separation using prestored spectra,” in *Proc. ICA*, 2006, pp. 561-568.

[18] F. R. Bach and M. I. Jordan, “Blind one-microphone speech separation: A spectral learning approach”, in *Proc. NIPS*, 2005, pp. 65-72.

[19] T. Tolonen, “Methods for separation of harmonic sound sources using sinusoidal modeling,” presented at the AES 106*th* Convention, Munich, Germany, 1999.

[20] T. Virtanen and A. Klapuri, “Separation of harmonic sound sources using sinusoidal modeling,” in *Proc. ICASSP*, 2000.

[21] T. Virtanen, “Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint,” in *Proc. DAFx*, 2003.

[22] M. R. Every and J. E. Szymanski, “Separation of synchronous pitched notes by spectral filtering of harmonics,” *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 5, pp. 1845-1856, 2006.

[23] Y. Li and D. L. Wang, “Separation of singing voice from music accompaniment for monaural recordings,” *IEEE Trans. Audio Speech Language Process.*, to be published.

[24] A. S. Bregman, *Auditory Scene Analysis*. The MIT Press, Cambridge, Massachusetts, 1990.

[25] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Comput. Speech Language*, vol. 8, pp. 297-336, 1994.

[26] B. Wang and M. D. Plumbley, “Musical audio stream separation by non-negative matrix factorization,” in *Proc. DMRN Summer Conference*, Glasgow, 2005.

[27] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio Speech Language Process.*, to be published.

[28] S. A. Abdallah and M. D. Plumbley, “Unsupervised analysis of polyphonic music using sparse coding,” *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 179-196, 2006.

[29] Y. Zhang and C. Zhang, ”Separation of music signals by harmonic structure modeling,” in *Proc. NIPS*, 2006, pp. 1617–1624.

[30] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, “Application of Bayesian probability network to music scene analysis,” in *Working Notes of IJCAI Workshop on CASA*, 1995, pp. 52-59.

[31] M. A. Casey and A. Westner, “Separation of mixed audio sources by independent subspace analysis,” in *Proc. ICMC*, 2000, pp. 154-161.

[32] C. Uhle, C. Dittmar, and T. sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” in *Proc. ICA*, 2003, pp. 843-848.

[33] M. K. I. Molla and K. Hirose, “Single-mixture audio source separation by subspace decomposistion of hilbert spectrum,” *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 3, pp. 893-900, 2007.

[34] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegatie matrix factorization,” *Nature*, vol. 401, pp. 788-791, 1999.

[35] M. Kim and S. Choi, “Monaural music source separation: nonnegativity, sparseness, and shift-invariance,” in *ICA*, 2006, pp. 617-624.

[36] R. Weiss and D. Ellis, “Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking,” in *Proc. Workshop Statistical Perceptual Audition (SAPA’06)*, Oct. 2006, pp. 31-36.

[37] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, “Model-based audio source separation,” Tech. Rep. C4DM-TR-05-01, Queen Mary University of London, 2006.

[38] J. Marques and P. Moreno, “A study of musical instrument classification using Gaussian mixture models and support vector machines,” *Cambridge Research Laboratory Technical Report Series CRL/4*, 1999.

[39] D. E. Hall, *Musical Acoustics, 3rd ed.* California State University, Sacramento, Brooks Cole, 2002.

[40] T. Kitahara, M. Goto, H.G. Okuno, “Pitch-dependent musical instrument identification and its application to musical sound ontology,” *Developments in Applied Artificial Intelligence*, Springer, 2003.

[41] V. Välimäki, J. Pakarinen, C. Erkut and M. Karjalainen, “Discrete-time modelling of musical instruments,” *Reports on progress in physics*, vol. 69, pp. 1-78, 2006.

[42] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[43] J. Eggink and G. J. Brown, “Application of missing feature theory to the recognition of musical instruments in polyphonic audio,” in *ISMIR*, 2003.

[44] X. Serra, “Musical sound modeling with sinusoids plus noise,” in *Musical Signal Processing*, C. Roads, S. Popea, A. Picialli, and G. D. Poli, Eds. Swets & Zeitlinger Publishers, 1997.

[45] The University of Iowa Musical Instrument Samples. [Online]. Available: http://theremin.music.uiowa.edu/.

[46] Sound Quality Assessment Material (SQAM) [Online]. Available: http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/.

[47] X. Rodet, Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models. presented at *IEEE Time-Frequency and Time-Scale Workshop*, 1997 [Online]. Available: http://recherche.ircam.fr/equipes/analyse-synthese/listePublications/articlesRodet/TFTS97/TFTS97-ASP.ps

[48] J. O. Smith, X. Serra “PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation,” in *ICMC*, 1987.

[49] M. Davy, S. Godsill and J. Idier, “Bayesian analysis of polyphonic western tonal music,” *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498-2517, 2006.

[50] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[51] H. Thornburg, R. J. Leistikow and J. Berger, “Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1257-1272, 2007.

[52] R. J. Leistikow, H. Thornburg, J. O. Smith and J. Berger, “Bayesian identification of closely-spaced chords from single-frame STFT peaks,” in *Proc. 7th Int. Conf. Digital Audio Effects (DAFx-04)*, Naples, Italy, 2004, pp. 228-233.

[53] A. Klapuri, “Multiple fundamental frequency estimation by summing harmonic amplitudes,” *ISMIR*, 2006.

[54] Y. Zhang, C. Zhang, and S. Wang, “Clustering in knowledge embedded space,” in *ECML*, 2003, pp. 480-491.

[55] M. Slaney, D. Naar, and R. F. Lyon, “Auditory model inversion for sound separation,” *ICASSP*, 1994, pp. 77-80.

[56] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 4, pp. 1462-1469, 2006.

[57] C. Févotte, R. Gribonval, and E. Vincent, *BSS EVAL Toolbox User Guide*, IRISA Technical Report 1706, Rennes, France, April 2005. http://www.irisa.fr/metiss/bss eval/.

[58] E. Vincent and R. Gribonval, *BSS ORACLE Toolbox User Guide Version 1.0*, 2005, URL: http://www.irisa.fr/metiss/bss oracle/

[59] T. Eerola, P. Toiviainen, *MIDI Toolbox: MATLAB Tools for Music Research*, University of Jyväskylä: Kopijyvä, Jyväskylä, Finland, 2004. http://www.jyu.fi/musica/miditoolbox/