Reconstructing individual monophonic instruments from musical mixtures using scene completion

Jinyu Han and Bryan Pardo
Northwestern University

2010

Abstract

Monaural sound source separation is the process of separating sound sources from a single channel mixture. In mixtures of pitched musical instruments, the problem of overlapping harmonics poses a significant challenge to source separation and reconstruction. One standard method to resolve overlapped harmonics is based on the assumption that harmonics of the same source have correlated amplitude envelopes: common amplitude modulation (CAM). Based on CAM, overlapped harmonics are approximated using the amplitude envelope from the nonoverlapped harmonics of the same note. CAM assumes nonoverlapped harmonics from the same noteare available and have similar amplitude envelopes to the overlapped harmonics. This is not always the case. A technique is proposed for harmonic temporal envelope estimation based on the idea of scene completion. The system learns the harmonic envelope for each instruments notes from the nonoverlapped harmonics of other notes played by that instrument, wherever they occur in the recording. This model is used to reconstruct the overlapped harmonic envelopes for obstructed harmonics. This allows reconstruction of completely overlapped notes, yet does not require predetermined instrument models. Experiments show the proposed algorithm performs better than an existing system based on CAM when the harmonics of pitched instrument are strongly overlapped.

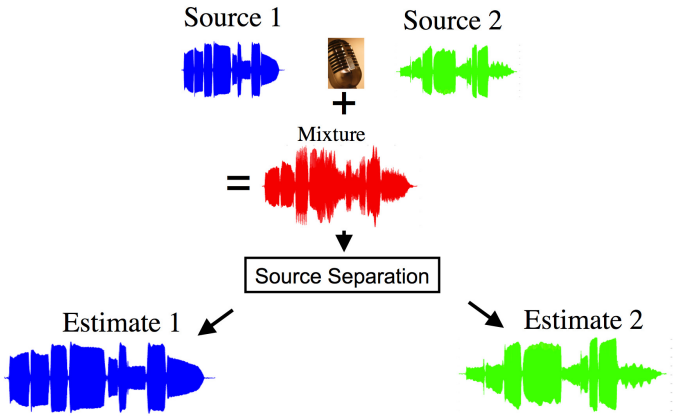
Contents

Chapter 1

Introduction

Musical sound separation is the process of isolating individual sound from a polyphonic mixture (e.g. singing voice and its accompaniment, a wind ensemble etc). Fig. 1.1 illustrates the process of single channel mixture recording and separation process. A solution to this problem has potential applications in many music information retrieval tasks, such as music transcription, content-based analysis, query by example system, and speech enhancement. Source separation would also facilitate post production of preexisting recordings, sample-based musical composition, multichannel expansion of mono and stereo recordings, and structured audio coding.

Figure 1.1: Single channel mixture recording and separation process



In this paper, we address the problem of monaural source separation of harmonic sounds, where multiple monophonic sounds produced by harmonic instruments are mixed to a single channel. The main contribution of this paper is that we proposed a new method for estimating the overlapped harmonics from a completely overlapped note.

The following sections define the problem of monaural harmonic sound separation and describe the related work in this area. We introduce the

background knowledge used in musical harmonic sound separation in Chapter 2. In Chapter 3 we present a new source separation approach, designed to isolate multiple simultaneous instruments from a single channel mixture of tonal music. The proposed method incorporates an existing technique based on sinusoidal model and improves on it by solving the completely overlapped harmonics problem that arises very often in recordings of tonal music. Chapter 4 provides a comparison of our algorithm to an existing source separation algorithm on real recordings of harmonic instruments, and a discussion of the advantages and limitations of using our approach. Finally, in Chapter 5 we summarize our findings and discuss directions for future research.

1.1 Problem Definition

When different sounds are recorded by a single microphone or mixed to a single channel, the observed time-domain signal is the linear superposition of individual source signals:

$$z(t) = \sum_{i=1}^{I} x_i(t) \tag{1.1}$$

where $x_i(t)$ is the signal of source i and $z(t)$ is the mixture. I is the number of sound sources. The task of monaural source separation is to isolate one or more source signals $x_i(t)$ from $z(t)$. Since the number of mixtures is less than the number of sources, the separation problem is underspecified. Knowledge at some level about the sources has to be assumed in order to solve this problem. In this paper, we assume the sound sources are monophonic harmonic sounds produced by musical instruments. 'Monophonic' means each source only has one fundamental frequency and 'harmonic' means the signal typically contains strong energy at integer multiples of its fundamental frequency called harmonics.

1.2 Related Work

Broadly speaking, existing monaural sound separation systems applied to music mixtures are either based on traditional signal processing techniques (mainly *sinusoidal modeling*) [1, 2], *computational auditory scene analysis* [3], or *statistical methods* [4, 5, 6] such as independent subspace analysis, sparse coding and nonnegative matrix factorization. The method proposed in this paper belongs to the first two categories, assuming the pitch track of each underlying source is already known. The estimation of pitch tracks [7, 8] in a polyphonic mixture is another important research problem in music information retrieval community.

Sinusoidal Modeling

Sinusoidal modeling [9, 10] assumes a sound can be represented by a linear combination of sinusoids with time-varying frequency, amplitude, and phase parameters. Fundamental frequencies of the harmonic sounds are often utilized to assist the source separation process. For a harmonic sound, each harmonic within a short period of time (compared to the rate-of-change of the sound) is represented as a sinusoid with fixed frequency, amplitude and phase parameters. So the task of the sound separation method is to estimate these parameters for each harmonic of each sound source in the mixture [2, 11, 12, 13]. Fundamental frequencies of each source are often used to identify the overlapped harmonics and estimate the frequency parameters of the sinusoids. A sinusoidal model provides a compact representation for harmonic sources. Thus sinusoidal model has often been used for separating harmonic sounds from a mixture. This paper has also adopted sinusoidal model to represent the harmonic sound sources. More details about sinusoidal model are discussed in Sec. 2.1.

Computational Auditory Scene Analysis

Computational auditory scene analysis (CASA) [3] is inspired by *auditory scene analysis (ASA)* [14], a perceptual theory that attempts to explain the remarkable capability of human auditory system to perform selective attention. Many CASA researchers try to create a symbolic representation of a sound scene in terms of individual sources[15]. Generally, CASA systems have two stages: segmentation (analysis) and grouping (synthesis). In segmentation, the acoustic input is decomposed into sensory segments, each of which originates from a single source. In grouping, the segments that likely come from the same source are put together.

The core of many monaural harmonic sound separation systems based on CASA[16][17][18][19] is a time-frequency (T-F) mask. Specifically, the Time-Frequency units in the acoustic mixture are selectively weighted in order to enhance the desired signal. The weights can be binary or real. The binary T-F masks are motivated by the masking phenomenon in human audition, in which a weaker signal is masked by a strong one in the same critical band [20]. Additionally, from the speech segregation perspective, the notion of an ideal binary mask has been proposed as the computational goal of CASA [21]. Such a mask can be constructed from a priori knowledge about target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference and 0 indicates otherwise.

Another very widely used time-frequency mask is one that assumes energy should be present in the harmonics of each harmonic source. In many sound separation systems [2, 22] inspired by CASA, a harmonic mask is constructed from the fundamental frequencies of each source to help the

source separation. By assuming that the input signals contain its main energy at its harmonics (integer multiples of its fundamental frequency), we could estimate the energy in regions where sources are not overlapped in the time-frequency representation by estimating each source's fundamental frequency and making harmonic masks that represent the expected high-energy time-frequency frames for each source. Due to the complexity and difficulty of multiple-pitch tracking in polyphonic music, the fundamental frequency of each source is usually assumed to be given or obtained partly by human correction. In this paper, harmonic mask is also utilized to help identify the overlapped and non-overlapped harmonics of each source. We concentrate on the separation of overlapped harmonics itself so the ground truth fundamental frequency of each source is assumed to be given in this paper. Our previous work on multiple-pitch tracking is described in [8, 23] and has reached promising results.

Statistical methods

Statistical methods for musical sound separation generally assume certain statistical properties of sound sources. Independent Component Analysis (ICA) [24, 25] assumes source signals are statistically independent, it iteratively determines time-invariant demixing filters to achieve maximal independence between sources. Independent subspace analysis (ISA) [26] extends ICA to single-channel source separation.

Sparse coding [6] assumes that source is a weighted sum of bases from an over-complete set. The weights are assumed to be mostly zeros, i.e., most of the bass are inactive most of the time.

Nonnegative matrix factorization (NMF) [5] attempts to find a mixing matrix and a source matrix with non-negative elements such that the reconstruction error is minimized. It implicitly requires the mixing weights are sparse[4].

The method proposed in this paper utilized techniques from the first two categories. Incorporating some statistical learning methods is one of our future research directions.

1.3 Motivation and Contribution

The main motivation behind this work is the remarkable capability of the human auditory system to separate sounds originating from different sources. For example, a human listener can single out a singing voice despite the accompaniment or follow several instruments simultaneously. Although these tasks seem to be effortless to humans, they turn out to be very difficult for machines. A robust monaural separation system could enhance the understanding of how the human auditory system performs these tasks, which remains a mystery at the present time.

Almost all music separation systems have to deal with the overlapped harmonics problem. Two harmonics of different sources overlap in the Time-Frequency domain when their frequencies are the same or close. In music that favors the twelve-tone equal temperament scale, a large number of harmonics of a given source may be overlapped by the harmonics of another source in the mixture. Resolving overlapping harmonics is the key to successfully reconstructing the original music sources.

Early separation systems based on CASA [16, 27] allocate energy in each time-frequency bin exclusively to one source and make no attempt to separate overlapping harmonics. Therefore the separation performance for these systems is limited.

The statistical methods handle overlapping harmonics implicitly, relying on the observed magnitudes in overlapped T-F regions to recover individual harmonic while ignoring the relative phases of the harmonic, which play a critical role in the observed magnitude spectrum. For example, assume that two overlapping harmonics have the same frequency and peak amplitude. If the relative phase of these two harmonics is 0, then the observed peak magnitude will be two times of the individual peak amplitude. However, if the relative phase is π, then the observed magnitude would be 0 because these two signal cancelled each other out. So the observed magnitude spectrum in the overlapped region will be different depending on the relative phase; thus, the phase information must be considered in order to accurately recover individual harmonic from the overlapped regions.

Recent systems that attempt to resolve the overlapped harmonics explicitly can be divided into two categories.

The systems inspired by CASA try to utilize the information of the neighboring non-overlapped harmonics to get reliable estimation of the overlapped region. Several different assumptions on the relationship of neighboring harmonics have been proposed. *Spectral smoothness* [28] assumes that the spectral envelope of instrument sound is smooth. Based on this assumption, the amplitude of an overlapped harmonic is estimated from the amplitudes of the neighboring non-overlapped harmonics using different kind of interpolation (Linear or Nonlinear) or weighting techniques[28, 29, 30]. Another assumption, known as *Common amplitude modulation (CAM)* [3] assumes that the amplitude envelopes of different harmonics of the same source tend to be similar so that the amplitude envelope of the overlapped harmonic could be approximated by the amplitude envelope of the non-overlapped harmonics of the same source. CAM has been utilized recently both for stereo and monaural musical sound separation and achieved good results.[31, 2].

Another way to deal with the overlapping harmonics is to use instrument model [32] that contain the relative amplitudes of harmonics. However, instrument-model based methods are limited because harmonic amplitude relationships are not consistent between recordings of different pitches, playing styles, and even different builds of the same instrument type.

The above-mentioned methods both failed when the energy of the available non-overlapped harmonics are too week or there is no non-overlapped harmonic available in the overlap region, e.g., one instrument playing one octave higher than the other one. Tonal music makes extensive use of multiple simultaneous instruments, playing consonant intervals. It is very common that the pitches of different instrument have integer relationship with each other; in which case, reliable non-overlapped harmonics of the higher pitched sound are not available. When two instruments are playing pitches of integer relationships, most harmonics of the higher pitched instrument are overlapped by the harmonics of the lower pitched instrument. We say the higher pitched instrument is "completely overlapped" in the mixture. The above-mentioned methods described in this section all failed to deal with the "complete overlap". In this paper, we proposed a novel framework to solve the "completely overlapped" harmonic amplitude estimation problem based on ideas from *Scene Completion.*

Scene Completion

Scene Completion [33] is the process of patching up missing sections in images by matching color and texture to other photos. It has been an active research area for years in the image processing community. This process has similarities with the above mentioned instrument-model method. Methods based on instrument models use the harmonic structure of similar notes to reconstruct the overlapped notes from the mixture while *Scene Completion* uses the content of other texture-similar images to patch the holes in the target image. However, the image produced with a scene completion method may be a totally different image, while the goal in sound separation is to recover the original underlying sound as accurately as possible. Due to inconsistency of harmonic amplitude relationships among different recordings, it is hard to reconstruct the underlying harmonic structure based on a harmonic structure model created from a different recording.

In a mixture containing several instruments playing simultaneously, we learn a linear model of the temporal harmonic envelope for each instrument from the non-overlapped reliable harmonics of that source throughout the recording. To separate a completely overlapped note from the mixture, this model is applied to reconstruct the harmonic envelope for each overlapped harmonic. Our proposed method incorporates the advantages of *Common Amplitude Modulation* but also allows dealing with completely overlapped notes. Since the envelope models are learned within the same recording as the overlapped notes are, it partly overcomes the limitation of the instrument model based method that the instrument model is inconsistent with the target notes to be separated.

Chapter 2

Background

2.1 Sinusoidal Model

A sinusoid[34] is any function of the form $\alpha \sin(\omega t + \phi)$,where α, ω, and ϕ are fixed amplitude, frequency, and phase parameters of the sinusoid respectively. Any tonal sound can be naturally and efficiently modeled as a sum of sinusoids over short period of time (compared to the rate-of-change of the sound). Over longer time durations, tonal sounds are well modeled by modulated sinusoids, where the amplitude and frequency parameters change slowly over time. Sinusoidal modeling [9, 10] is a well established technique in audio synthesis and signal processing. It models a sound sources as the summation of individual sinusoidal components.

In Sec. 1.1, we defined the general problem of musical sound separation. In this section, we will show how to use sinusoids to model tonal sounds. Given a harmonic sound source, we break it into small segments of short period of time (e.g., several tens of milliseconds), called frames. Within an analysis frame with index m, Eq. 1.1 can be rewritten in the discrete time domain.

$$z^m[n] = \sum_{i=1}^{I} x_i^m[n] \tag{2.1}$$

where m and n denotes the frame index and sample index within the frame respectively.

Since we assume harmonic sound sources, each sound source can be expressed as a sum of sinusoids at frequencies given by integer multiples of its fundamental frequency. The harmonics of each source could be characterized as time-variant sinusoids. Within an analysis frame of suitable length, the frequencies and amplitudes of the sinusoids can be assumed constant. The sinusoidal model of a harmonic sound $x_i^m[n]$ (source i at the nth sample of the mth frame) can be written as

$$x_i^m[n] = \sum_{h_i=1}^{H_i} \alpha_i^{h_i}(m) \cos(2\pi f_i^{h_i}(m) n T_n + \phi_i^{h_i}(m)) \tag{2.2}$$

where $\alpha_i^{h_i}(m)$ and $f_i^{h_i}(m)$ are the amplitude and frequency parameters respectively, of the h_i ($h_i = 1, ..., H_i$) harmonic of source i within time frame m. $\phi_i^{h_i}(m)$ is the phase of h_i harmonic of source i at the beginning of time frame m. H_i denotes the number of harmonics in source i and T_n denotes the sampling period in seconds.

The sinusoidal model of $x_i^{(m)}[n]$ can be transformed to the time-frequency domain by the discrete Fourier transform (DFT) using an analysis window $w[n]$. The DFT of $x_i^{(m)}[n]$, windowed by $w[n]$, at frequency bin k is

$$X_i(m,k) = \sum_{h_i=1}^{H} \frac{\alpha_i^{h_i}(m)}{2} (e^{j\phi_i^{h_i}(m)} W(kf_b - f^{h_i}(m)) \tag{2.3}$$

$$+ e^{-j\phi_i^{h_i}(m)} W(kf_b + f^{h_i}(m))) \tag{2.4}$$

where $f_b = f_s/N$ is the frequency resolution of the DFT, f_s is the sampling frequency. W is the discrete-time Fourier transform (DTFT) of the analysis window of the same length as the frame in this paper:

$$W(f) = \sum_{n=0}^{N-1} w[n] e^{-j2\pi(f/f_s)n} \tag{2.5}$$

where N is the length of the DFT.

For a perfectly harmonic sound, $f^{h^i}(m) = h^i F_i(m)$, where $F_i(m)$ denotes the fundamental frequency of source i at time frame m, if we assume that $W(f) \approx 0$ for $|f| > \theta_1$, where θ_1 is a threshold in Hz, then $|W(kf_b + f^{h^i}(m))| \approx 0$ provided $F_i(m) > \theta_1$ at time frame m. Furthermore, if $F_i(m) > 2\theta_1$, then $|W(kf_b - f^{h^i}(m))| > 0$ for at most one harmonic of source i, allowing us to drop the summation over harmonics from Eq. 2.3. This means the harmonics of the same source are not overlapped with each other in the Time-Frequency domain given a suitable analysis window of the DFT. Given the above assumptions, the DFT of $x_i^{(m)}[n]$ in Eq. 2.3 could be further simplified as:

$$X_i(m,k) = \frac{\alpha_i^{h_i}(m)}{2} e^{i\phi_i^{h_i}(m)} W(kf_b - h_i F_i(m)) \tag{2.6}$$

where F_i denotes the fundamental frequency of source i at time frame m.

Assuming that the mixing process is linear as illustrated in Eq.2.1, the sinusoidal model of a mixture of I harmonic sound sources in the time-

frquency domain can be written as

$$Z(m,k) = \sum_{i=1}^{I} X_i(m,k). \tag{2.7}$$

This model treats a polyphonic mixture as a collection of harmonic components from multiple sound sources. Given the fundamental frequency F_i of each source i, the task of musical sound separation is to estimate $\{\alpha_i^{h_i}(m), \phi_i^{h_i}(m)\}$ for all the harmonic components of the I sources.

As shown in Eq.2.2, the phase change of a harmonic is related to the instantaneous frequency of a sinusoid as follows:

$$\phi_i^{h_i}(m+1) - \phi_i^{h_i}(m) = 2\pi f_i^{h_i}(m) T_m \tag{2.8}$$

The above equation is equivalent to

$$\triangle\phi_i^{h_i}(m) = 2\pi f_i^{h_i} T_m = 2\pi h_i F_i(m) T_m \tag{2.9}$$

Here T_m denotes the hop size of the STFT in seconds. The relationship gives us the progression of a harmonic's phase from the sources' fundamental frequencies, provided the signal adheres to the harmonic sinusoidal model, the frequency is stable over the duration of the time frame and the pitch estimate is accurate.

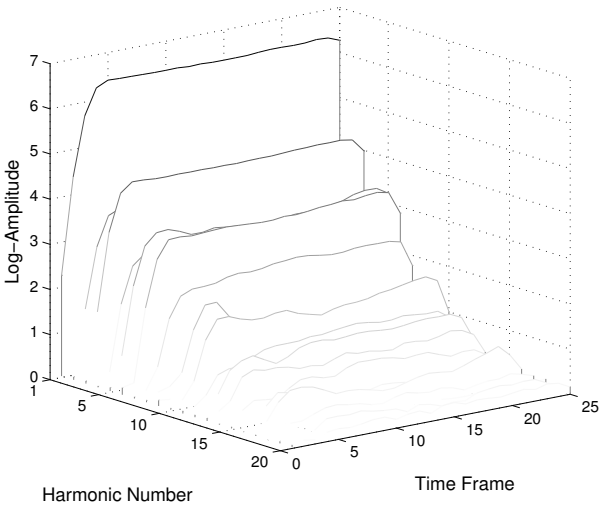


Figure 2.1: Logarithm of the amplitude envelopes for the first 20 harmonics of a clarinet playing a F4

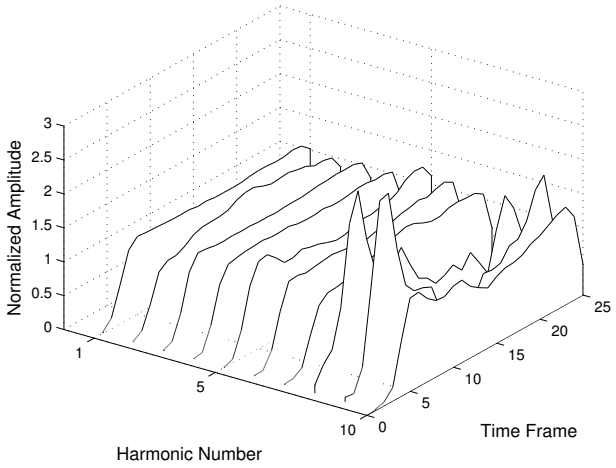


Figure 2.2: The normalized amplitude envelopes for the first 10 harmonics, which contain 98% energy of the same F4 note played by a clarinet

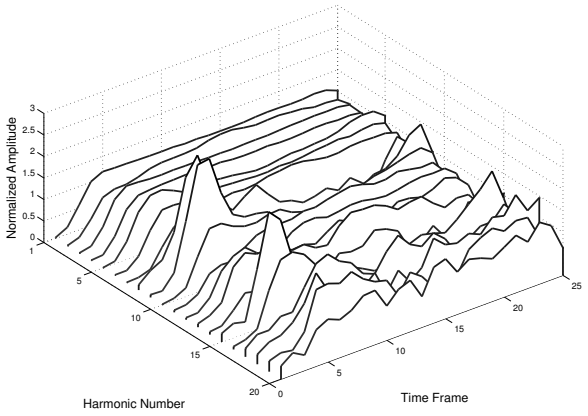


Figure 2.3: First 20 harmonic amplitude envelopes normalized by the average amplitude of each harmonic of the same F4 note played by a clarinet

2.2 Common Amplitude Modulation

Common Amplitude Modulation (**CAM**) assumes that the amplitude envelopes of spectral components from the same sound source are correlated. Fig. 2.1 showed the amplitude envelopes of first 20 harmonics of a clarinet playing the pitch F4. It suggests that, although the amplitudes of different harmonics are quite different, the envelopes of the harmonics, especially the strongest ones, do share the same general modulation trend.

The amplitude envelopes of the first 10 harmonics, which consist 98% of

the energy from the same note from Fig. 2.1 are plotted in Fig. 2.2, where each harmonic is normalized by the its average amplitude. Comparing it to Fig. 2.3, where the normalized envelopes of the first 20 harmonics are plotted, we can see that *CAM* holds most of the time for the first a few harmonics with strong energy, while fails to hold for those with week energy.

In [2], the correlation coefficient between the strongest harmonic of an individual instrument tone with the other harmonics is calculated as a function of difference in amplitude. The box plots of the results taken from [2] is shown in Fig. 2.4. We cam see that the correlation is high for harmonics with energy close to that of the strongest harmonic and tapers off as the energy in the harmonic decreases. The evidence from [2] agrees with our assumption that *CAM* holds most of the time for harmonics with strong energy.

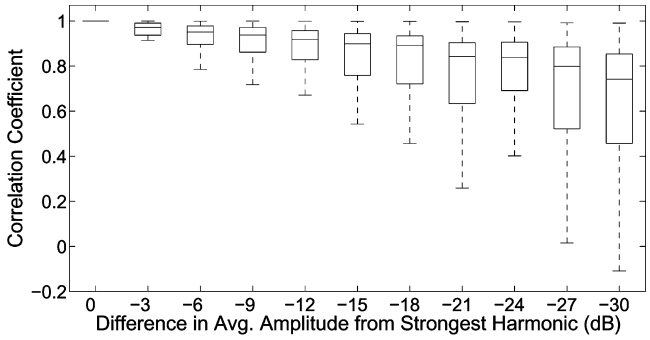


Figure 2.4: Box plots of correlation coefficients measured between the strongest harmonic and other harmonics of individual instrument notes for the sustained portions of each note, plotted as a function of amplitude difference between harmonics. Results are calculated using 100 note samples. From [2], by permission of the authors

Since the low-energy harmonics do not have a strong influence on the perception of a signal and high-energy harmonics follow the *CAM*, we could impose the harmonic amplitude envelope of high-energy harmonic on all of the harmonics and not change the perception of a tone too much.

Fig. 2.5 and 2.6 in next page show an example of imposing *CAM* on all harmonics of two tones played by bassoon and violin. The common harmonic envelope is taken from the harmonic with strongest energy of each tone and the amplitude value of the first frame for each harmonic is estimated so as to minimizes the difference between the original harmonic amplitude and regenerated harmonic amplitude. A small human subject study [1] performed by the author showed that the regenerated tones using *CAM* is perceptually

[1]This study was participated by five students and has never been published

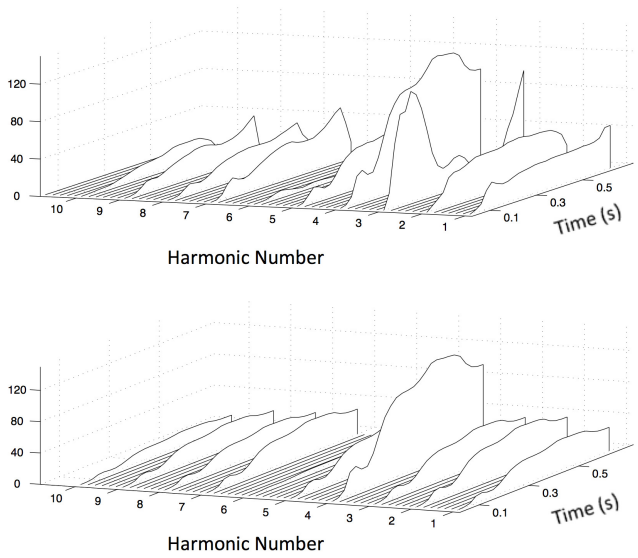


Figure 2.5: The figure above is the amplitude of the first 10 harmonics of the original note played by a bassoon; The figure below is the amplitude of the first 10 harmonics of the re-synthesized note by imposing *CAM*

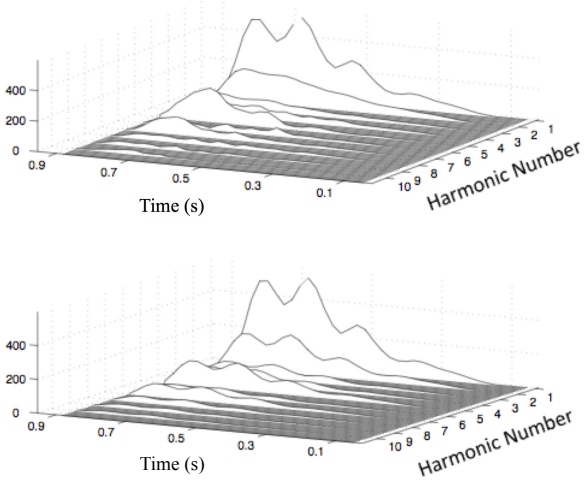


Figure 2.6: The figure above is the amplitude of the first 10 harmonics of the original note played by violin; The figure below is the amplitude of the first 10 harmonics of the re-synthesized note by imposing *CAM*

indistinguishable from the original tones.

The above-mentioned empirical evidences suggest that the amplitude envelope of an overlapped harmonic could be approximated from the amplitude envelopes of non-overlapped harmonics of the same source within the same note, provided that the non-overlapped harmonics have strong enough energy.

2.3 Harmonic Temporal Envelope Similarity

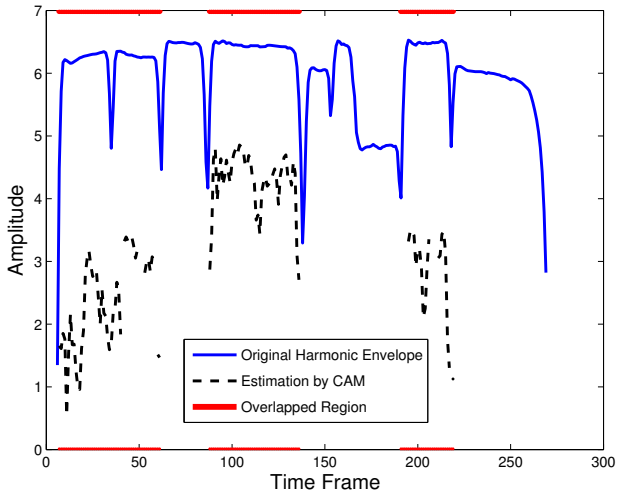


Figure 2.7: Comparison between the Original Harmonic envelope and the estimated envelope by CAM. The original envelopes of the first harmonic from 9 notes played by clarinet are plotted in solid blue line. The red line indicates the overlapping region. Four notes played a clarinet with fundamental frequencies 398.4Hz, 397.7Hz, 296.6Hz and 293.3Hz are completely overlapped with four notes with fundamental frequencies 132.6Hz, 198.2Hz, 98.2Hz and 146.9Hz, played by bassoon. The dashed black line is the estimated envelope by CAM. The available non-overlapped harmonic for these four overlapped notes have harmonic numbers 48, 48, 39 and 39

In this section, we show empirical evidence for harmonic envelope approximation based on information from non-overlapped harmonics of other notes.

Tonal music makes extensive use of multiple simultaneous instruments, playing consonant intervals. It is very common that the pitches of different instrument have integer relationship with each other, in which case, most harmonics with strong energy from the higher pitched instrument are overlapped, leaving the non-overlapped harmonics with very high harmonic

numbers. Since most harmonics with high harmonic numbers have very weak energy, they do not follow the same amplitude modulation as the high-energy harmonics do.

Fig.2.7 showed an example of harmonic envelope estimation by *CAM* when almost all harmonics of the overlapped notes are overlapped with another lower-pitched instrument. It is clear that the harmonic envelope of the first harmonic has little similarity with the harmonic envelope estimated by *CAM* from the 39th or 48th harmonics. A close examination showed that the harmonic envelopes of these four overlapped notes have very similar shape with the envelopes from the non-overlapped notes (e.g. the 3rd note in this example). Fig.3.5 from Sec. 3.2 of Page. 25 showed the harmonic envelope estimation of the same four overlapped notes based on the envelope information from the 3rd and last non-overlapped notes.

Although the harmonic temporal amplitude evolution of tones played by different instruments may be very different, the harmonic temporal envelope of different notes played by the same instrument within a short period of time usually shows great resemblance.

Fig. 2.8 and Fig. 2.9 showed the first 10 harmonic envelopes of four consecutive notes played by a clarinet and bassoon. We can see that although these four notes have different fundamental frequencies and lengths, their amplitude envelopes of the strong-energy harmonics evolve similarly to each other.

This similarity of harmonic envelope among different notes played by the same instrument exists commonly in wind instruments. For string instrument, the similarity among different notes is less obvious which is supported by the experiment results presented in Chap. 4.

This empirical evidence shows that the envelopes of strong-energy harmonics from different notes of the same instrument are usually better correlated with each other than the envelopes of harmonics of strong-energy and weak-energy from the same note. We will show that in Sec.3.2 how to use the harmonic envelope of an non-overlapped note to approximate the changes of harmonic amplitude of an overlapped note.

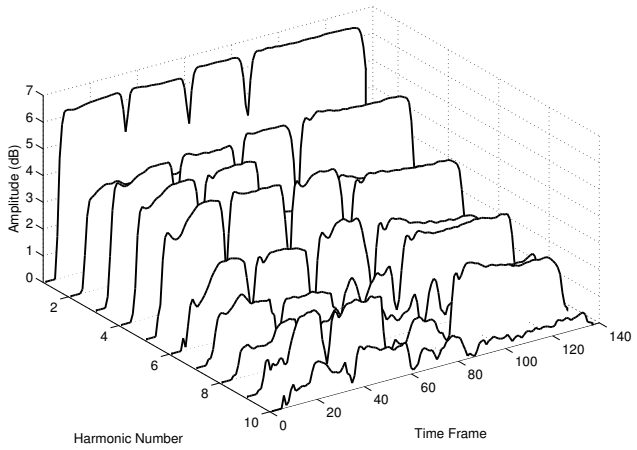


Figure 2.8: The first 10 harmonics of four consecutive notes played by a clarinet within a same piece. The pitches of the four notes are 400Hz, 375Hz, 300Hz and 330Hz

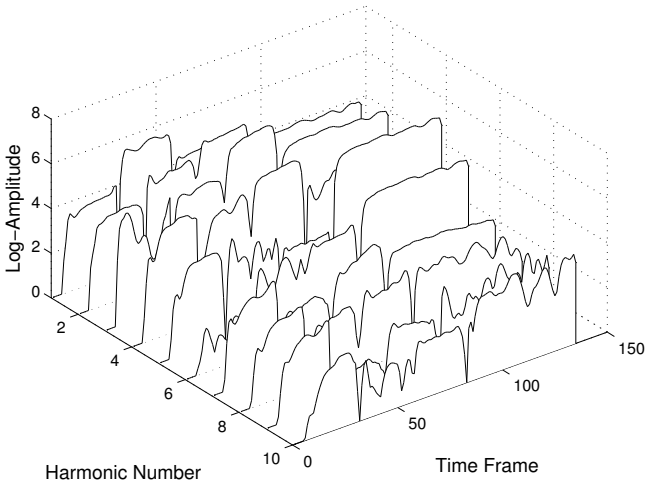


Figure 2.9: The first 10 harmonics of four consecutive notes played by a bassoon within a same piece. The pitches of the four notes are 132Hz, 147Hz, 197Hz and 100Hz

Chapter 3

Method Description

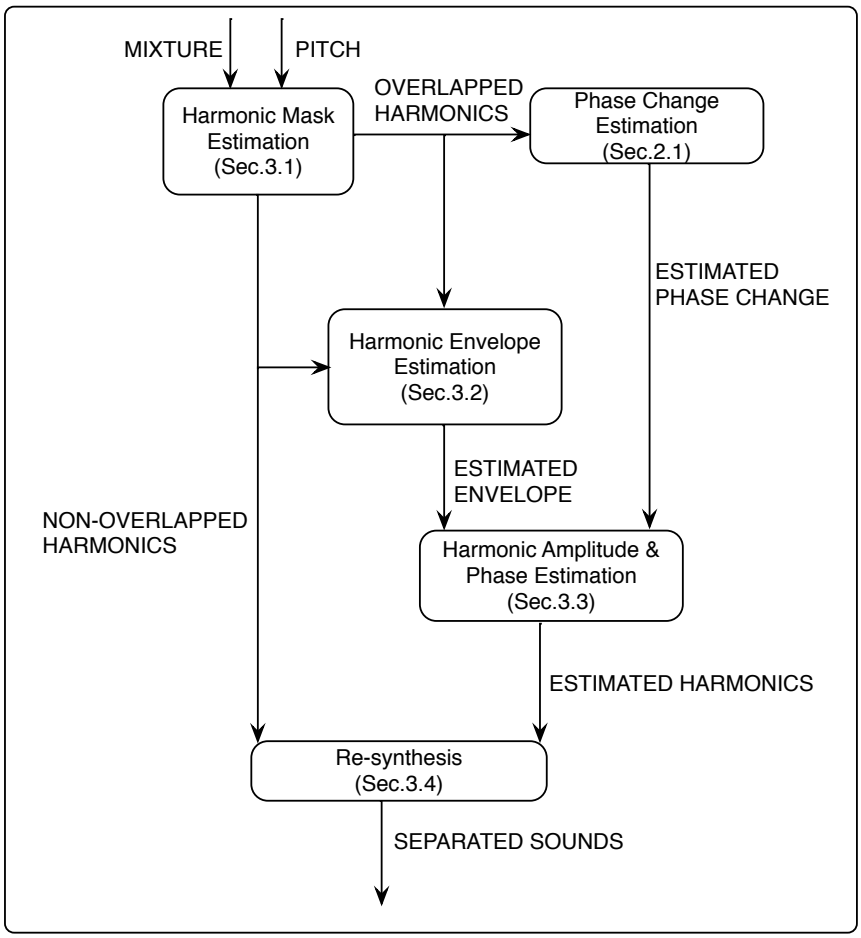


Figure 3.1: System overview. Rectangle indicates the functional module in our system. Arrow indicates the input and output of the functional module

Our proposed separation system is illustrated in Fig. 3.1. The input

to the system is a polyphonic, single-channel mixture and fundamental frequency of each source. We address the source separation problem in four stages. The first stage is *Harmonic Mask Estimation* (Sec. 3.1), where the input fundamental frequencies are used to construct a harmonic mask for each source to identify the non-overlapped and overlapped harmonics. In the second stage, *Harmonic Envelope Estimation* (Sec. 3.2), the harmonic envelopes of the non-overlapped harmonics are modeled by a linear function model and the harmonic envelopes of the overlapped harmonics are estimated by these linear models. In the stage of *Harmonic Phase and Amplitude Estimation* (Sec. 3.3), the amplitudes and phases for the overlapped harmonics from different sources are estimated. Given the phase changes estimated from the instantaneous frequencies and harmonic envelopes of the overlapped harmonics estimated from the second stage, the initial phase and amplitude of each overlapped harmonic are estimated under a least-square estimation framework. The resulting amplitude and phase parameters are used to estimate the STFT values for each source in the overlapped T-F regions and these values are passed to the *Re-synthesis* (Sec. 3.4) stage and added to the STFT values from the non-overlapped bins identified by the harmonic masks. Finally, the overlap-add method is used to convert the estimated STFT of each signal to a time-domain estimate of each source.

The main contribution of this paper lies in the second stage. **We proposed a new framework to solve the problem of estimating the overlapped harmonic envelope when the source is completed overlapped with other sources.**

We propose a simple but efficient method to reconstruct overlapped harmonic envelopes using the envelopes of the non-overlapped harmonics wherever they are available. It is based on the same idea as the scene completion technique in computer vision that the corrupted sections in images could be patched up using sections of other photos with similar color and textures. In our framework, the corrupted sections correspond to the overlapped harmonics and our goal is to find a similar harmonic envelops for the overlapped harmonics from the harmonics that are not overlapped. We utilized the property that notes played by the same instrument within a short period of time have similar harmonic envelopes. When the non-overlapped harmonics of the same note is available and reliable, we use the linear model built from the non-overlapped harmonic of the same note to estimate the envelope of the overlapped harmonic. In this case, our approach is equivalent to using *CAM* for harmonic envelope estimation. When reliable non-overlapped harmonics are not available, we utilize the linear model built from the non-overlapped harmonics of other notes that have similar length to the target harmonic. The experiment results showed that using the harmonic envelope of similar note, we could achieve relatively reliable envelope estimation and better separation of the overlapped harmonics.

3.1 Harmonic Mask Estimation

The first processing stage takes as input a polyphonic mixture signal and pitch estimates for each source signal. This stage first transforms the input using STFT into spectrogram and use the fundamental frequency estimates to generate a harmonic mask for each source by identifying the frequency bins associated with each harmonic at each time frame. A frequency bin k at time frame m is associated with harmonic h_i of source i if

$$|kf_b - f_i^{h_i}(m)| < \theta_1 \tag{3.1}$$

where θ_1 is a same threshold described in Sec.2.1.

We denote the set of frequency bins associated with h_i of source i at frame m as $K_i^{h_i}(m)$:

$$K_i^{h_i}(m) = \{k \mid |kf_b - f_i^{h_i}(m)| < \theta_1\} \tag{3.2}$$

We can define overlapped and non-overlapped harmonics similarly. Harmonic h_i of source i is overlapped by some other harmonics h_j of source j at time frame m if

$$|f_i^{h_i}(m) - f_j^{h_j}(m)| < \theta_2 \tag{3.3}$$

where θ_2 is also a threshold.

In this case, we say harmonic h_i from source i is overlapped with harmonic h_i from source j. Notice that when $\theta_1 > \theta_2/2$, it means that a frequency bin could be assigned to multiple harmonics of different sources.

If no other harmonic has a frequency within θ_2 of harmonic h_i, we call h_i non-overlapped and denote the set of non-overlapped harmonics for source i in frame m as $\tilde{H}_i(m)$. Furthermore, we say the set of frquency bins $K_i^{h_i}(m)$ associated with harmonic h_i of source i at time frame m is non-overlapped if h_i itself is non-overlapped:

$$|f_i^{h_i}(m) - f_j^{h_j}(m)| \geq \theta_2, \forall j \neq i, \forall h_j\} \tag{3.4}$$

A harmonic mask is simply a collection of overlapped harmonics at each time frame and their associated frequency bins. We construct $M_i(k, m)$, the harmonic mask for source i, by finding each frequency bin k at time frame m, that belongs to a overlapped harmonic of source i. We place a 1 in each of these elements in the harmonic mask as shown in Eq.3.5

$$M_i(k,m) = \begin{cases} 1 & if\ k \in K_i^{h_i}(m)\ \&\ h_i\ is\ overlapped \\ 0 & o.w. \end{cases} \tag{3.5}$$

Given the harmonic masks for each source, the non-overlapped bins of each source are identified by Eq.3.6.

$$\tilde{X}_i(k,m) = Z(k,m) \times (1 - M_i(k,m)) \tag{3.6}$$

where $\tilde{X}_i(k,m)$ denote all the non-overlapped bins of source i.

The non-overlapped bins of each source are passed to the *Re-synthesis* stage directly. After the overlapped and non-overlapped harmonics are identified for each source, the non-overlapped harmonics are used to construct the linear model for harmonic envelope in Sec. 3.2 and the overlapped harmonics are further processed in Sec.3.2 and Sec.3.3.

3.2 Harmonic Envelope Estimation

In Sec.2.2 and Sec.2.3, we presented empirical evidences on using the non-overlapped harmonic envelope to approximate the overlapped harmonic envelope. In this stage, we describe the details of our proposed framework on harmonic envelope estimation based on "Scene Completion"

Note Model Construction

The amplitudes from a non-overlapped harmonics h_i of source i identified in Sec. 3.1 are estimated by finding the amplitude $\alpha^{h_i}(m)$ that minimizes Eq.3.7 in all the bins $k \in K_i^{h_i}(m)$:

$$\sum_{k \in K_i^{h_i}(m)} (|Z(m,k)| - \frac{\alpha^{h_i}(m)}{2}|W(kf_b - h_i F_i(m))|)^2 \tag{3.7}$$

where $|Z(m,k)|$ is the observed amplitude of the spectrogram of frequency bin k at time frame m.

The minimization of the above equation is

$$\alpha^{h_i}(m) = \frac{2\sum_{k \in K^{h_i}(m)} |Z(k,m)| \cdot |W(kf_b - h_i F_i(m))|}{\sum_{k \in K^{h_i}(m)} |W(kf_b - h_i F_i(m))|^2} \tag{3.8}$$

where $W(kf_b - h_i F_i(m))$ is calculated using Eq. 2.5. This gives us an estimation of the amplitude parameter for the non-overlapped harmonics of each source.

For one single note, let $t \equiv (m_1, ..., m_N)^T$ denote the time frame indices associated with it, and $r \equiv (r_1, ..., r_N)^T$ denote the corresponding normalized harmonic envelope where $r_l = \alpha^{h_i}(m_l)/\alpha^{h_i}(m_1)$ estimated using Eq.3.8. Here, h_i is the available non-overlapped harmonics with strongest energy from the same note. We re-index the frame indeces by $(x_1, ..., x_N)^T = (1, ..., N)^T$. Fig.3.2 shows a plot of a normalized harmonic envelope of a note with length $N = 24$. The envelope is obtained by estimating the harmonic amplitude of the first harmonic of a note played by a clarinet and normalized by the amplitude of its first frame.

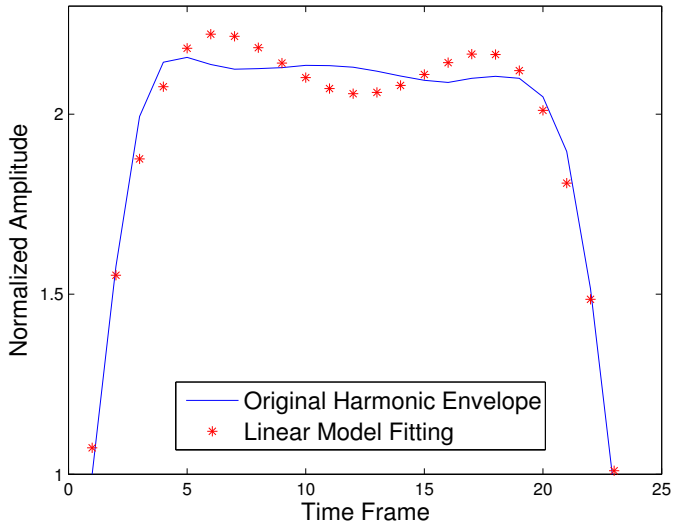


Figure 3.2: The original harmonic envelope of a note played by clarinet and its linear model: $y = 0.4019 + 0.7804 \times x - 0.1158 \times x^2 + 0.0069 \times x^3 - 0.0001 \times x^4$

Our goal is to exploit the harmonic envelopes from the non-overlapped harmonics to make predictions for the envelopes of the overlapped harmonics. In this paper, we consider this as a curve-fitting problem and fit the envelope data using a polynomial function of the form:

$$y(x, w) = \omega_0 + \omega_1 x + \omega_2 x^2 + ... + \omega_M x^M = \sum_{j=0}^{M} \omega_j \theta_j(x) \tag{3.9}$$

where $y(x, w)$ is the predicted envelope value at time index x. M is the order of the polynomial, $\theta_j(x) = x^j$ is the *basis function* and x^j denotes x raised to the power of j. The polynomial coefficients $\omega_0, ..., \omega_M$ are collectively denoted by the vector w.

The values of the coefficients will be determined by fitting the polynomial to the harmonic envelope. This can be done by minimizing an *error function* that measures the misfit between the function $y(x, w)$, for any given value of w, and the training set data points which is the observed non-overlapped harmonic envelope. One simple widely used *error function* described in Eq.3.10 is given by the sum of the squares of the errors between the predictions $y(x_l, w)$ for each time index x_l and the corresponding target values r_l.

$$E(w) = \frac{1}{2} \sum_{l=1}^{N} \{y(x_l, w) - r_l\}^2 \tag{3.10}$$

The solution w^* minimizing Eq. 3.10 is obtained by:

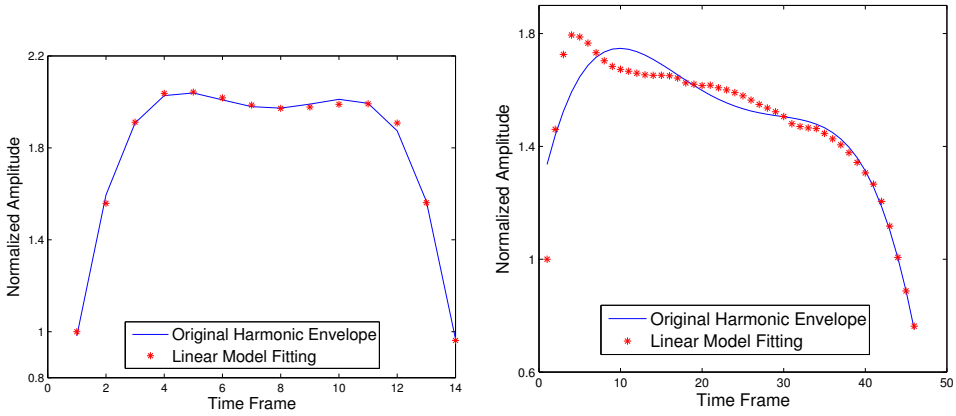
$$w^* = (\theta^T \theta)^{-1} \theta^T \mathbf{r} \tag{3.11}$$

where θ is an $N \times M$ matrix, whose elements are given by $\theta_{lj} = \theta_j(x_l)$:

$$\theta = \begin{pmatrix} \theta_0(x_1) & \theta_1(x_1) & ... & \theta_{M-1}(x_1) \\ \theta_0(x_2) & \theta_1(x_2) & ... & \theta_{M-1}(x_2) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \theta_0(x_N) & \theta_1(x_N) & ... & \theta_{M-1}(x_N) \end{pmatrix} \tag{3.12}$$

For every note which is not completely overlapped, we construct a linear model (w^*, N) for it, where $w*$ is the polynomial coefficients and M is the length of the note. In Fig. 3.2, we showed an example of the result of fitting polynomial having order $M = 5$ to a harmonic envelope. Fig.3.2 showed more linear model fitting results to different kind notes played by a clarinet.

Figure 3.3: Linear Model fitting of order 5 to two notes played by a clarinet with different lengths and shape dynamics



Given a completely overlapped note of length L, we could approximate its harmonic envelope r^* using an existing linear model (w^*, N) learned from another non-overlapped harmonics of length N by Eq.3.13

$$r^* = \omega_0^* + \omega_1^* x + \omega_2^* x^2 + ... + \omega_M^* x^M = \sum_{j=0}^{M} \omega_j^* \theta_j(x) \tag{3.13}$$

where $x_i = 1 + (i-1) \times \frac{N-1}{L-1}$ for $i = 1, ..., L$.

Envelope Estimation

Given a overlapped harmonic, our proposed method for estimating the harmonic envelope is illustrated in Fig. 3.4. When a note is note "completely overlapped" by other sources, we use the linear model built from the envelope

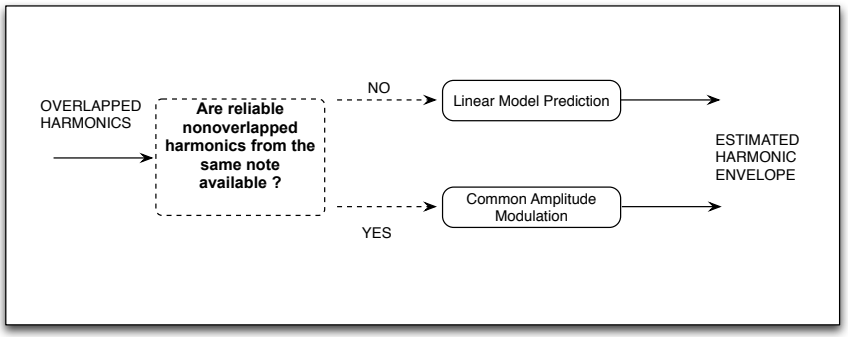


Figure 3.4: Framework for overlapped harmonic envelope estimation

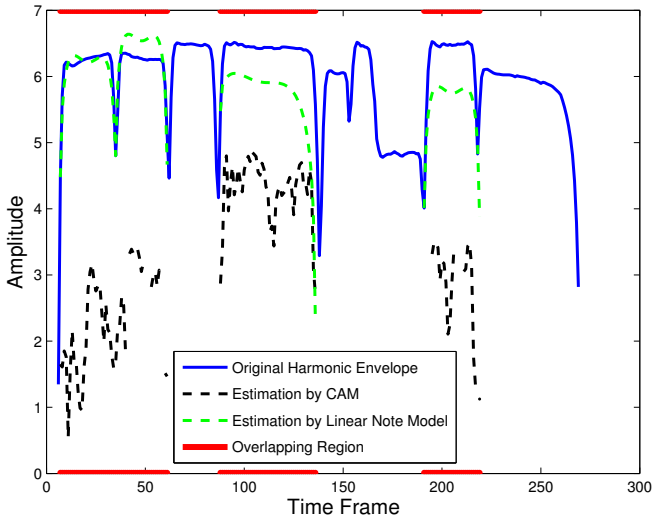


Figure 3.5: Comparison between the Original Harmonic envelope and the estimated envelope by *Linear Model*. The original envelopes of the first harmonic from 9 notes played by a clarinet are plotted in solid blue line. The red line indicates the overlapping region. Four notes with fundamental frequencies 398.4Hz, 397.7Hz, 296.6Hz and 293.3Hz are completely overlapped with four notes played by a clarinet with fundamental frequencies 132.6Hz, 198.2Hz, 98.2Hz and 146.9Hz, played by a bassoon. The dashed black line is the estimated envelope by *CAM*. The dashed green line is the estimated envelope by *Linear Model*. The available non-overlapped harmonic for these four overlapped notes have harmonic numbers 48, 48, 39 and 39

of the strongest non-overlapped harmonic of the same note to approximate the overlapped harmonics. Otherwise, we find another note that has the closest length to the length of the target note, and use the linear model learned from that note to regenerate a new envelope by Eq.3.13. These re-

estimated envelopes are used in the next stage of our system to estimate the initial amplitude value of the overlapped harmonics.

Fig.3.5 showed the estimated harmonic envelope by *Linear Model* described in this section for the same completely overlapped notes from Sec. 2.3. The first, second, fourth and eighth notes are completely overlapped by another instrument playing lower pitches. The first available non-overlapped harmonics for these four notes have harmonic number 48, 48, 39 and 39 respectively. The estimation based on the non-overlapped harmonics of the same by *CAM* are illustrated in dashed blacked line, which are very unstable and different from the original envelope. The dashed green line is the envelope estimated by utilizing the linear model built from the third note and last note of the example. It showed that our proposed model produces much better envelope estimates for the completely overlapped notes than the *CAM* does.

The estimated harmonic envelopes, along with the estimated phase changes by Eq.2.9, are used in next stage to estimate the amplitude and phase of the overlapped harmonics under a least-square estimation framework.

3.3 Harmonic Phase and Amplitude Estimation

Given an overlapped harmonic in a sequence of continuous time frame, the phase change of this harmonic could be estimated using the sinusoidal model described in Eq. 2.9 from Sec. 2.1, and the envelope of this harmonic could be estimated based on the framework described in Sec.3.2. The parameters remained to be estimated are the initial phase and amplitude of the overlapped harmonic. In this section, we described a method to estimate the initial phase and amplitude under a least-square estimation framework.

The hop size of the STFT is in the tens of milliseconds, which tends to be shorter than the length of individual notes. As a result, overlap between harmonics often occurs in sequences of time frames as well as a series of frequency bins. Accordingly, we extend the idea of overlapped harmonic to an overlapped T-F region. Let $\{h_{i_1}, ..., h_{i_P}\}$ be a set of P harmonics from sources from i_1 through i_P that overlap during time frames from m_0 to m_1. The overlapped T-F region for this set of harmonics is defined as

$$D(m_0, m_1; k_0, k_1) = \{m, k | m \in \{m_0 ... m_1\}; k \in \{k_0 ... k_1\}\} \tag{3.14}$$

where k_0 is the smallest $k \in U_{i=i_1}^{i_P} U_{m=m_0}^{m_1} K_i^{h_i}(m)$, k_1 is the largest and $K_i^{h_i}(m)$ denotes the set of frequency bins associated with h_i of source i at frame m.

The above-defined overlapped region is the bounding box that includes the frequency bins associated with all of the overlapping harmonics. For example, assume that h_{i_1} and h_{i_2} overlap during time frames 10 through 18

during frequency bins 21 through 26. Then the overlapped T-F region is $D(10, 18; 21, 26)$.

According to Eq. 2.6 and Eq.2.7 from Sec.2.1 the observed STFT value $Z(m, k)$ of the mixture can be written as

$$Z(m,k) = \sum_i S_i^{h_i}(m) W(kf_b - h_i F_i(m)) \tag{3.15}$$

where

$$S_i^{h_i}(m) = \frac{\alpha_i^{h_i}(m)}{2} e^{i\phi_i^{h_i}(m)} \tag{3.16}$$

is the sinusoidal parameter of harmnic h_i from source i.

The relation between $S_i^{h_i}(m)$, the sinusoidal parameter in current frame m, and $S_i^{h_i}(m_0)$, the sinusoidal parameter in the initial frame m_0 of the bounding box is as follow:

$$S_i^{h_i}(m) = S_i^{h_i}(m_0)(\gamma_m^{h_i})(e^{i\sum_{\iota=m_0}^{m} \triangle\phi_n^{h_i}(\iota)}) \tag{3.17}$$

where $\gamma_m^{h_i}$ is the estimated envelope value of harmonic h_i at frame index m and $\triangle\phi_n^{h_i}(\iota)$ is the phase change during the frame ι calculated by Eq.2.9. The envelope of harmonic h_i was estimated from the previous section and normalized so that the initial value $\gamma_{m_0}^{h_i} = 1$.

We could rewrite Eq. 3.15 as:

$$Z(m,k) = \sum_i S_i^{h_i}(m_0) R_i^{h_i}(m,k) \tag{3.18}$$

where $R_i(m, k)$ is defined as follow:

$$R_i^{h_i}(m,k) = W(kf_b - h_i F_i(m))\gamma_m^{h_i}(e^{i\sum_{\iota=m_0}^{m} \triangle\phi_n^{h_i}(\iota)}) \tag{3.19}$$

In Eq.3.18, $Z(m, k)$ is the observed DTFT value of the mixture and $R_i^{h_i}(m, k)$ only depends on the harmonic envelope and phase changes of the harmonics, both of which are estimated from previous sections. $S_i^{h_i}(m_0)$, the initial amplitude and phase of harmonic h_i, is the only term to be estimated.

We further rewrite Eq. 3.18 in overlapping region $D(m_0, m_i; k_0, k_i)$ in matrix format:

$$\begin{pmatrix} R_1(m_0,k_0) & ... & R_N(m_0,k_0) \\ ... & ... & ... \\ R_1(m_i,k_i) & ... & R_N(m_i,k_i) \end{pmatrix} \begin{pmatrix} S_1^{h_1}(m_0) \\ ... \\ S_N^{h_i}(m_0) \end{pmatrix} = \begin{pmatrix} X(m_0,k_0) \\ ... \\ X(m_i,k_i) \end{pmatrix} \tag{3.20}$$

$$\Downarrow\Downarrow$$

$$RS = X \tag{3.21}$$

The least-squares estimation of S is given by:

$$S = (R^H R)^{-1} R^H X \tag{3.22}$$

Where H denotes the conjugate transpose.

After S is estimated, the complex sinusoidal parameter of each harmonic at all the frames contributing to the overlapped region can be estimated using Eq.3.17.

3.4 Re-synthesis

In the final estimation of the STFT of each source signal, we combine the spectrogram from the non-overlapped harmonics and the estimates from the overlapped harmonic regions.

In Sec.3.1, we have already shown how to use Eq.3.6 to generate the spectrogram $\tilde{X}_i$ from the non-overlapped region of source i. For the bins associated with overlapped harmonics h_i, we utilize the sinusoidal model to calculate the STFT as follows:

$$\hat{X}_i(m,k) = S_i^{h_i}(m) W(k f_b - h_i F_i(m)) \tag{3.23}$$

Finally, the overall source STFT is

$$X_i = \hat{X}_i + \tilde{X}_i \tag{3.24}$$

and we use the overlap-add method to obtain the time domain estimate $x_i[n]$ for each source i.

Chapter 4

Experiment

4.1 Dataset and Experiment Setup

The proposed system was tested on a dataset extracted from 10 real music performances, totaling about 330 seconds of audio. Each performance was of a four-part Bach chorale, performed by a quartet of instruments: violin (soprano), clarinet or trumpet (alto), tenor saxophone (tenor) and bassoon (bass). Each musician's part was recorded in isolation while the musician listened to the others through headphones.

We call a note "completely overlapped" in some period of time if almost all of its harmonics are overlapped with harmonics from another instrument. This happens when a instrument is playing a pitch of roughly integer number of the pitch from another instrument. For example, if source i and source j are playing simultaneously with fundamental frequencies $398.2Hz$ and $132.6Hz$ respectively, and θ_2 is set to be $16.14Hz$, according to the Eq. 3.3, the first 40 harmonics of source i are all overlapped with harmonics of source j. The first available non-overlapped harmonics is thus harmonic 41, which has very low energy and shows very unstable harmonic envelope. In this case, we say source i is "completely overlapped" within the note of $398.2Hz$ fundamental frequency. Our algorithm is designed to separate these "completely overlapped" harmonics while the previous methods all failed.

We tested our algorithm on mixtures of two instruments with one instrument (bassoon) playing the bass line and the other playing the alto (clarinet or trumpet) or soprano line (violin) of the Bach chorale mentioned above. Since we are only interested in the separation results of the higher-pitched instruments where the "completely overlap" happens very often, the separation results on three instruments (violin, clarinet and trumpet) playing higher pitches are reported in this paper because they are extensively "completely overlapped" by the bass line. The instruments playing the bass or tenor line are not "completely overlapped" and can be separated very well

Table 4.1: Segments of mixtures at musical phrase boundaries

Mixtures [a]	Number of segments	Ave. length	Total length
Clarinet (alto)	50	6.67s	333.71s
Truempet (alto)	48	6.63s	324.34s
Violin (soprano)	49	6.95s	340.75s

[a]Bassoon as the bass line

Table 4.2: Completely overlapped notes for clarinet, trumpet and violin

Mixtures [a]	Number of notes	Ave. length	Total length
Clarinet (alto)	207	1.11s	230s
Truempet (alto)	214	1.07s	229s
Violin (soprano)	228	1.14s	260s

[a]Bassoon as the bass line

by just using *CAM*.

The mixtures we tested on contain two instruments. They are bassoon and clarinet, bassoon and trumpet, or bassoon and clarinet. For the convenience of experiment setup, each performance was segmented at its musical phrase boundaries into segments of roughly 5 to 8 seconds in length. Averagely, about two third of the segments are "completely overlapped". More information about the segments of different instrument was listed in Table. 4.1.

In order to show the performance result only on the "completely overlapped" notes, we further segmented segments in Table. 4.1 into smaller note-level segments which only contains two overlapped notes from two instruments, one of which is completely overlapped by the other. This procedure produced more than 200 completely overlapped notes for each instrument. The statistics of these completely overlapped notes are listed in Table 4.2.

It should be noticed that we only run the experiment on the segments at musical phrase boundaries, but not on the note-level segments. This is because our method needs non-overlapped harmonics to build the linear note models for each instrument. The performance on the completely overlapped notes is obtained by comparing the separated notes from the segments to the original notes.

In mixing, all signals are mixed with equal energy. In testing, the audio is broken into frames with length of 93 ms and 23 ms hop. No zero-padding is used in the DFT.

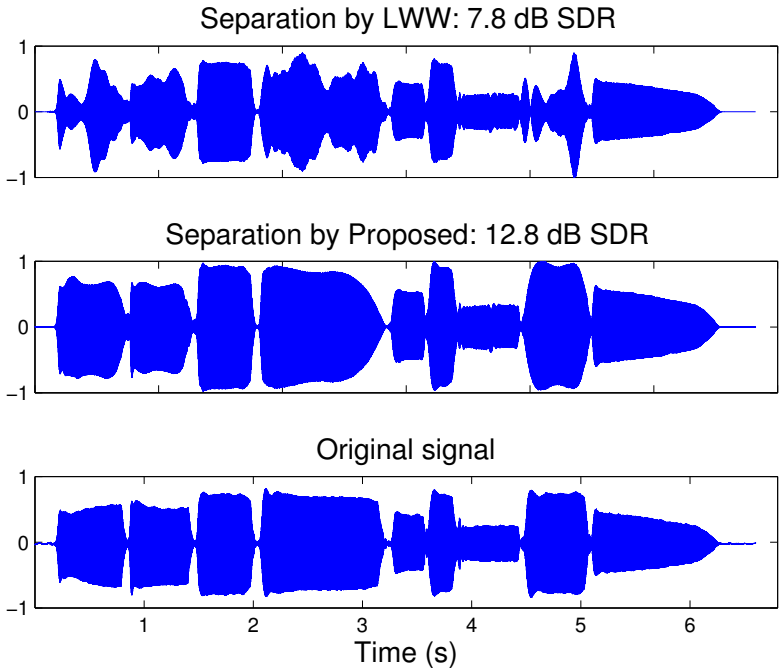
Hamming window was used in the DFT. We set θ_1 using the magnitude spectrum of the windowing function W. We associate frequency bins with a

harmonic until the magnitude windowing function W has dropped by $40dB$. We choose $\theta_2 = 1.5f_b$, which is approximately the 6-dB bandwidth of the Hamming window. The number of harmonics for each source H_i is chosen such that $f_i^{H_i}(m) < f_s/2$ for all time frames, where f_s denotes the sampling frequency. The sampling rate of all recordings is 44.1 kHz.

As mentioned earlier, the input to our system is the polyphonic mixture and the fundamental frequency of individual source. The ground-truth fundamental frequencies of each testing piece were estimated using [35] on monophonic sound tracks prior to mixing.

4.2 Experiment Results

Figure 4.1: Separation example of a clarinet from a 6.5 seconds mixture of clarinet and bassoon. SDR measurement showed there was a 5 dB improvement of the proposed method over LWW



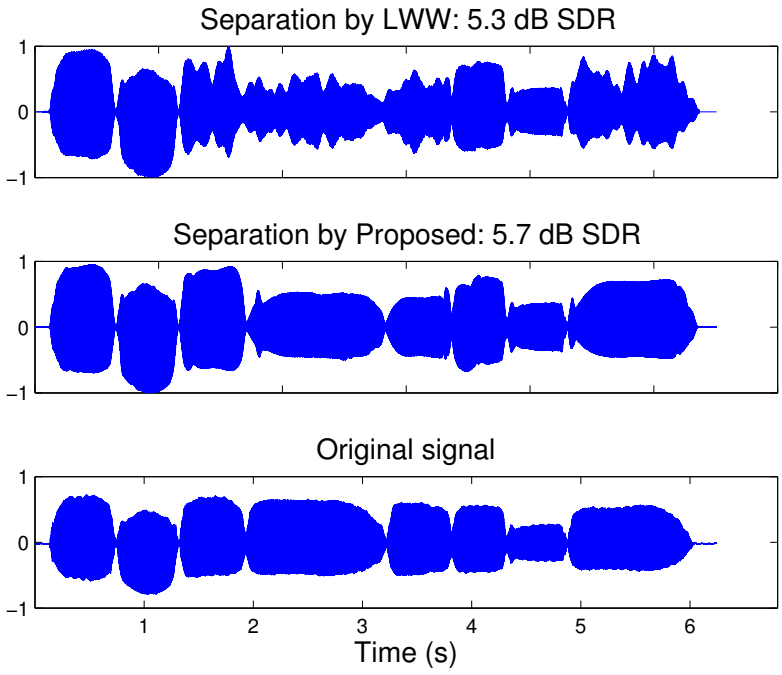
We compare the proposed system to a recent musical separation system[2], denoted by "LWW". "LWW" is a state-of-art harmonic musical sound separation system based on CASA and Sinusoidal Model. It exploits the assumption *CAM* that the harmonics of the same source have correlated amplitude envelopes.

The difference between the proposed system and "LWW" lies in the esti-

mation of the overlapped harmonic envelopes. "LWW" is based on *CAM*, using the non-overlapped harmonics of the same note to predict the envelopes of the overlapped harmonics. This approach gets more problematic when the non-overlapped harmonics has very low energy, or does not work when the non-overlapped harmonics of the same note are not available. These two cases happen very often when two instruments are playing pitches with integer relationship with each other. Our proposed method is designed to resolve this problem by grabbing the note envelope of another note to help predict the note envelope of the currently "completely overlapped" notes.

Separation Examples

Figure 4.2: Separation examples of a trumpet from a 6 seconds mixture of trumpet and bassoon, our proposed method and LWW have the same SDR measurement on the separated signals but produce a perceptually better separated signal



In this section, we present some empirical evidences that our proposed method produces superior separation results than the "LWW" does on the completely overlapped notes by comparing the waveform of the separated signals to the waveform of the original signals.

Fig. 4.1 on Page 31 showed a real separation example of clarinet from a mixture segment of clarinet and bassoon. There is an improvement (SDR)

by 5 dB of the proposed method over "LWW". The waveform of the separated signal by "LWW" showed the "completely overlapped" notes (the first, second, fourth, seventh and eighth note) have very irregular envelope. Our proposed system successfully learned the harmonic envelope for clarinet from the non-overlapped harmonics of other notes and applied these learned models to the completely overlapped notes. Comparing the separated signal by our proposed system to the original signal, we could see that although the envelopes of the completely overlapped notes are somewhat different from their original envelopes, the regenerated envelopes preserve the main characteristics of the shape of "a clarinet note" in this segment. This creates a perceptually similar reconstruction of the overlapped notes using the "texture" from the non-overlapped notes.

Figure 4.3: Separation example of a clarinet from a 4.5 seconds mixture of clarinet and bassoon. The separation performance of our proposed method has decreased by 2 dB by the measure of SDR

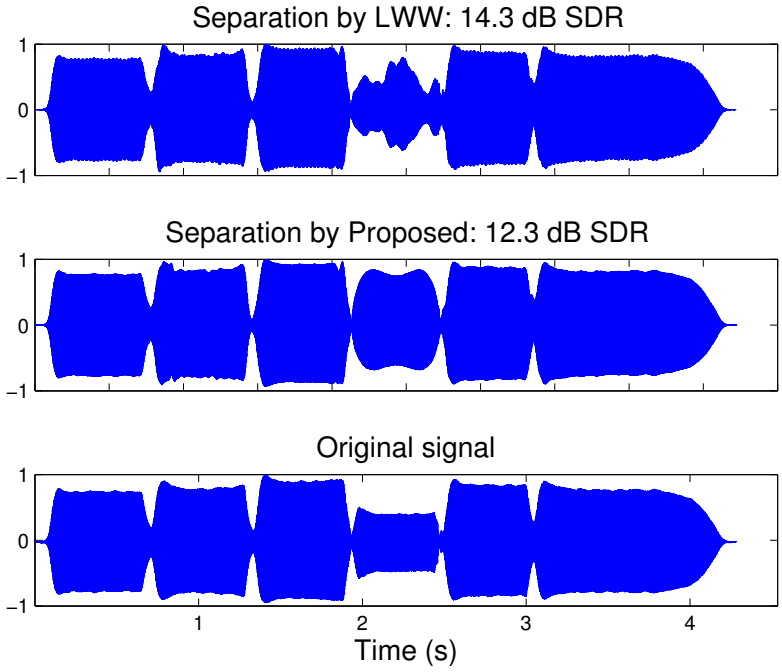


Fig 4.2 on Page 32 showed another example where there was no SDR improvement between our proposed method and "LWW", but the note envelopes of the completely overlapped notes separated by our proposed system are more similar to the original note envelopes played by trumpet in this segment, than the "LWW" does. Although there was no performance improvement measured by SDR, judging from the waveform of the separated signal our proposed method produced better separation results compared

the "LWW".

Fig. 4.3 further showed an example where there was performance decreasing of the proposed method compared "LWW". However, a close look at the waveform of the separated signals by "LWW" showed that the forth note of segment, which is "completely overlapped", has very irregular envelope shape. The envelope for this note by "LWW" is actually taken from the 50th harmonic that is the first non-overlapped harmonic in the same note. On the contrary, our proposed method applied the harmonic envelope model learned from the other non-overlapped notes, which produced a much better envelope estimation result. Although the separation performance was decreased by $2dB$ by the measure of SDR, it is clearly shown by the waveform of the separated signals that the separated notes by our proposed system have a better envelope shape than the "LWW" does.

Quantitative Result

Besides providing the real separated examples presented in previous section, we also measured the performance of our proposed method by some quantitative measurement. In this section, we describe the quantitative separation results of our experiments, and compare them to the separation results by "LWW".

The separation results are measured using source-to-distortion ratio (SDR), source-to-interfering ratio (SIR), and source-to-artifacts ratio (SAR) proposed in [36] for evaluation of sound separation algorithms. SDR, SIR and SAR measure overall distortion, energy from interfering sources and artifacts introduced by the separation algorithm, respectively. Results from preliminary study [37] indicate that these measures correlate more closely with human perception of signal similarity than other measures.

Overall separation performance on segments at musical phrase boundaries from Table. 4.1 are shown in Table 4.3 and Fig. 4.2. The proposed system improved the separation performance of Clarinet and Trumpet on SDR and SAR. Specifically, the average improvement on Clarinet is about 1.9 dB measured both by SDR and SAR, and 1.1 dB on Trumpet. *Student Test* showed that there are significant differences between the proposed method and LWW on performance measured by SDR and SAR but not on SIR. For the performance on violin, there is no significant difference between the proposed method and LWW. One reason is that violin has very unstable harmonic envelope and it is hard to characteristic the harmonic envelope using linear model. More complex model need to be applied to model the harmonic envelope of violin.

The separation results on the "completely overlapped" notes (the notes which are completely overlapped by another instrument) described in Table.4.2 are shown in Table. 4.4 and Fig. 4.2. The proposed system achieved a performance improvement on the completely overlapped notes of clarinet and

Table 4.3: Performance result on segments at musical phrase boundaries. Average SDR, SAR and SIR of the proposed system and the LWW are shown here. Results are obtained over 50 5-to-8 seconds long segments of totally about 330 seconds per instrument. Higher value of SDR, SAR and SIR means better separation result. Numbers in bold indicate that there are statistical differences on the measurement

Mixtures[a]	SDR		SAR		SIR	
	Proposed	LWW	Proposed	LWW	Proposed	LWW
Clarinet	**12.28**	10.41	**12.31**	10.42	42.29	41.85
Trumpet	**10.72**	9.61	**10.75**	9.62	41.38	39.71
Violin	6.33	6.37	6.34	6.38	44.12	43.68

[a]bassoon as the bass line

Table 4.4: Performance result on completely overlapped notes. Average SDR, SAR and SIR of the proposed system and the LWW are shown here. Results are obtained over more than 200 notes of totally more than 230-second long per instrument. Higher value of SDR, SAR and SIR means better separation result. Numbers in bold indicate that there are statistical differences on the measurement

Mixtures[a]	SDR		SAR		SIR	
	Proposed	LWW	Proposed	LWW	Proposed	LWW
Clarinet	**11.70**	9.68	**11.82**	9.79	37.73	34.65
Trumpet	**10.57**	9.43	**10.64**	9.49	39.21	38.42
Violin	6.04	6.37	6.12	6.41	38.00	38.55

[a]bassoon as the bass line

trumpet by 2 dB and 1.1 dB respectively. "Student T Test" showed that the separation performance of our proposed system measured by SDR and SAR are statistically better than "LWW" on clarinet and trumpet. There is on statistical difference on the separation results of violin.

Figure 4.4: Separation performance on clarinet signal from mixtures of two instruments. The separation results on segments at musical phrase boundaries are shown on the left. The separation results on note-level segments in Table.4.2 are shown on the right

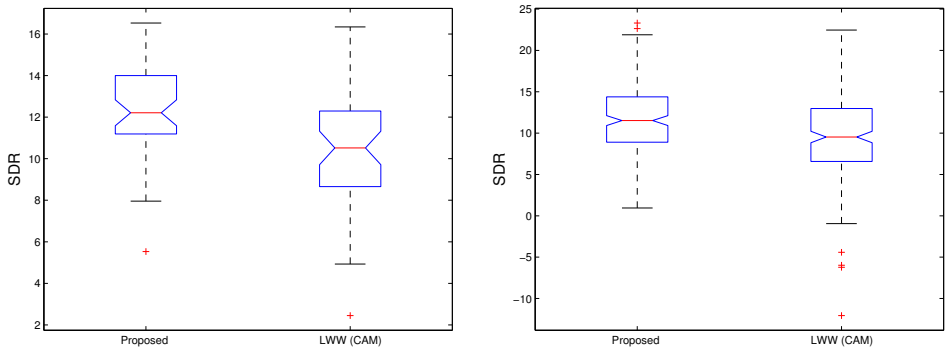


Figure 4.5: Separation performance on trumpet signal from mixtures of two instruments. The separation results on segments at musical phrase boundaries are shown on the left. The separation results on note-level segments in Table.4.2 are shown on the right

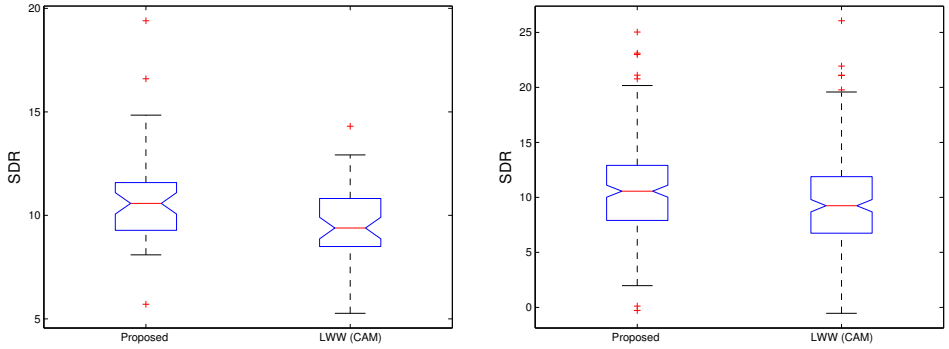
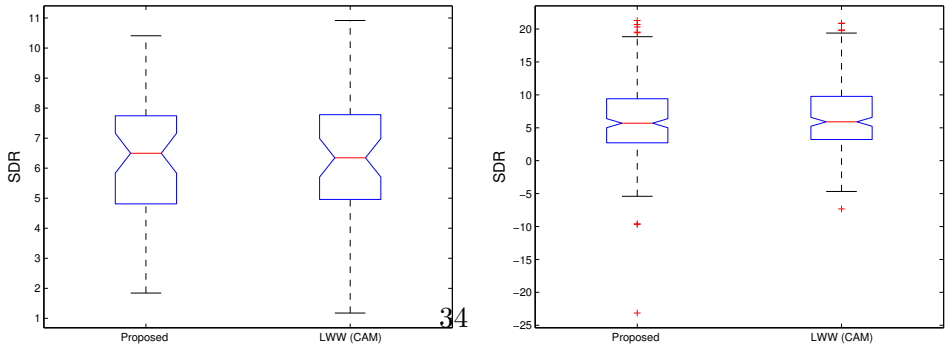


Figure 4.6: Separation performance on violin signal from mixtures of two instruments. The separation results on segments at musical phrase boundaries are shown on the left. The separation results on note-level segments in Table.4.2 are shown on the right



Chapter 5

Conclusion

In this paper, we proposed a monaural musical sound separation system that explicitly deals with the "completely overlapped" notes. Inspired by the idea "Scene Completion" from image processing, our approach is based on *Common Amplitude Modulation (CAM)* and the harmonic envelope similarity of different notes from the same instrument.

Quantitative results showed that when pitches can be estimated accurately, and the harmonic envelope of the instrument is stable among different notes, the separation performance achieves better separation performance than a state-of-art monaural music separation system that only exploits *CAM*. In addition to the improvement in quantitive measurement of SDR and SAR, the perceptual quality of the separated signals is improved judging by the waveform of the separated signals.

We have shown that the harmonic envelope of an instrument can be modeled using a linear function to some extend. The experiment results showed that the proposed linear model learned from the same recording of the instrument could be used to get stable prediction for harmonic envelope of the overlapped harmonics from another note, overcoming the disadvantages of instrument model. This approach works especially better for wind instrument that has a stable harmonic envelope. For instruments with unstable harmonic envelope such as violin, more sophisticated models need to be investigated to show the superiority of our method.

Bibliography

[1] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 2, pp. II765–II768 vol.2.

[2] Yipeng Li, John Woodruff, and DeLiang Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 7, pp. 1361–1371, 2009.

[3] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley, 2006.

[4] Daniel D. Lee and Sebastian H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[5] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[6] Paris Smaragdis, Madhusudana Shashanka, and Bhiksha Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I.Williams, and A. Culotta, Eds., pp. 1705–1713. 2009.

[7] A. Klapuri, "Multipitch analysis of polyphoneic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 255–266, 2008.

[8] Z.Duan, J.Han, and B.Pardo, "Harmonically informed pitch tracking," in *Proc. of the 10th International Conference on Music Information Retrieval*, 2009.

[9] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, Aug 1986.

[10] X Serra, "Musical sound modeling with sinusoids plus noise," *Musical Signal Processing*, pp. 497–510, 1997.

[11] M.Every and J.Szymanski, "A spectral-filtering approach to music signal separation," in *Proceedings of the 7th International Conference on Digital Audio Effects*, 2004.

[12] T Virtanen, *Audio signal modeling with sinusoids plus noise*, Master of science thesis, Tampere University of Technology, 2000.

[13] T Virtanen, *Sound Source Separation in Monaural Music Signals*, Phd thesis, Tampere University of Technology, 2006.

[14] A.S.Bregman, *Auditory Scene Analysis: The Perceptual Organization of sound*, The MIT Press, 1990.

[15] D.F.Rosenthal and H.G.Okuno, *Computational Auditory Scene Analysis*, Lawrence Erinbaum Associates, 1998.

[16] G.N. Hu and D.L. Wang, "Monaural speech separation," in *Advances in Neural Information Processing Systems 15*, S. Thrun S. Becker and K. Obermayer, Eds., pp. 1221–1228. MIT Press, Cambridge, MA, 2003.

[17] G.Hu and D.L.Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Neural Networks, IEEE Transactions on*, vol. 15, pp. 1135–1150, 2004.

[18] M. Wu and D.L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[19] N. Roman and D.L. Wang, "Pitch-based monaural segregation of reverberant speech," *The Journal of the Acoustical Society of America*, vol. 120, pp. 458–469, 2006.

[20] B.C.J. Moore, *An Introduction to the Psychology of Hearing, Fifth Edition*, Academic Press, 2003.

[21] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, pp. 181–197, 2005.

[22] J.Han and B.Pardo, "Improving separation of harmonic sources with iterative estimation of spatial cues," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

[23] Z.Duan, J.Han, and B.Pardo, "Song-level multi-pitch tracking by heavily constrained clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[24] A.Hyvärinen, "Survey on independent component analysis. neural computing surveys," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[25] L. Parra and C. Spence, "Separation of non-stationary natural signals," Independent Component Analysis: Principles and Practice, 135–157., 2001.

[26] Te-Won Lee G.J.Jang and Yung-Hwan Oh, "Single-channel signal separation using time-domain basis functions," *Signal Processing Letters, IEEE*, vol. 10, pp. 168–171, 2003.

[27] Yipeng Li and DeLiang Wang, "Musical sound separation using pitch-based labeling and binary time-frequency masking," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 173–176.

[28] A.P. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, vol. 5, pp. 3381–3384 vol.5.

[29] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[30] M.R. Every and J.E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1845–1856, Sept. 2006.

[31] J.Woodruff and B.Pardo, "Using pitch, amplitude modulation and spatial cues for separation of harmonic instruments from stereo music recordings," *EURASIP Journal on Applied Signal Processing*, vol. 2007, 2007.

[32] M. Bay and J. Beauchamp, "Harmonic source separation and polyphonic pitch detection using prestored spectra," in *6th International Conference on Independent Component Analysis and Blind Source Separation*, 2006.

[33] James Hays and Alexei A. Efros, "Scene completion using millions of photographs," *Communications of the ACM*, vol. 51, pp. 87–940, 2008.

[34] Julius O. Smith, *Spectral Audio Signal Processing, October 2008 Draft*, `//ccrma.stanford.edu/~jos/sasp/`, 2007, online book.

[35] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise of a sampled sound," in *Proc. the Institute of Phonetic Sciences*, 1993, vol. 17, pp. 97–110.

[36] E.Vincent, R.Gribonval, and C.Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 2006.

[37] B.Fox, A.Sabin, B.Pardo, and A.Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, 2007.