Chapter Number

Monaural Audio Separation Using Spectral Template and Isolated Note Information

Anil Lal and Wenwu Wang
Department of Electronic Engineering, University of Surrey, United Kingdom

1. Introduction

Musical sound separation systems attempt to separate individual musical sources from sound mixtures. The human auditory system gives us the extraordinary capability of identifying instruments being played (pitched and non-pitched) from a piece of music and also hearing the rhythm/melody of the individual instrument being played. This task appears 'automatic' to us but has proved to be very difficult to replicate in computational systems. Many methods have been developed recently for addressing this challenging source separation problem. They can be broadly classified into two categories, respectively, statistical learning based techniques such as independent component analysis (ICA) and non-negative matrix/tensor factorization (NMF/NTF), and computational auditory scene analysis (CASA) based techniques.

One of the popular methods for source separation was based on ICA [1-10], where the underlying unknown sources are assumed to be statistically independent, so that a criterion for measuring the statistical distance between the distribution of the sources can be formed and optimised either adaptively [5] [10] [11], or collectively (in block or batch processing mode) [2], given mixtures as the input signals. Both high-order statistics (HOS) [2] [4] [5] and second order statistics (SOS) [12] have been used for this purpose. The ICA techniques have been developed extensively since the pioneering contributions in early 1990s, made for example by Jutten [1], Comon [3], and Cardoso [2]. The early work of ICA concentrates on the instantaneous model which was soon found to be limited for real audio applications such as in a cocktail party environment, where the sound sources reach listeners (microphones) through multi-path propagations (with surface reflections). Convolutive ICA [13] was then proposed to deal with such situations (see [21] for a comprehensive survey). Using Fourier transform, the convolutive ICA problem can be converted to multiple instantaneous but complex valued ICA problems in the frequency domain [14-17] thanks to its computational efficiency, and the sources can be separated after permutation correction for all the frequency bins [18-20]. Most of the aforementioned methods consider (over-) determined cases where the number of sources is assumed to be no greater than the number of observed signals. In practical situations, however, an underdetermined separation problem is usually encountered. A widely used method for tackling this problem is based on sparse signal representations [22-29], where the sources are assumed to be sparse in either the time domain or a transform domain such that the overlap between the sources at

each time instant (or time-frequency point) is minimal. Audio signals (such as music and speech) become sparser when transformed into the time-frequency domain, therefore, using such a representation, each source within the mixture can be identified based on the probability of each time-frequency point of the mixture that is dominated by a particular source, using either sparse coding [26], or time-frequency masking [30] [31] [33], based on the evaluation of various cues from the mixtures, including e.g. statistical cues [20], and binaural cues [30] [32]. Other methods for source separation include, for instance, the non-negative ICA method [34], the independent vector analysis (IVA) [35], and NMF/NTF [37-43]. Comprehensive review of ICA (and other statistical learning) methods is out of the scope of this chapter, and for more references, we refer the interested readers to the recent handbook on ICA edited by Comon and Jutten [36], and a book on NMF by Cichocki [44].

Many ICA methods discussed above can be broadly applied to different types of signals. In contrast, CASA is another important technique dealing specifically with audio signals, which is based on the principles of Auditory Scene Analysis (ASA). In [60], Bregman attempts to explain ASA principles by illustrating the ability of the human auditory system to identify and perceptually isolate several sources from acoustic mixtures by separating the sources into individual (perceptual) acoustic streams for each source, which suggests that the auditory system operates in two main stages, segmentation and grouping. The segmentation stage separates the mixture into (time-frequency) components that would relate to an individual source. The grouping stage then groups the components that are likely to be from the same source e.g. using information such as simultaneous onset/offset of particular frequency amplitudes or relationships of particular frequencies to source pitch [45-50]. It is well-known that the ICA technique is not effective in separating the underdetermined mixtures, for which, as mentioned above, one has to turn to, e.g. the technique of sparse representations, by sparsifying the underdetermined mixtures into a transform domain, and to reconstruct the sources using sparse recovery algorithms [51-53]. In contrast, CASA technique evaluates the temporal and frequency information of the sources directly from the mixtures, and therefore it has the advantage in dealing with underdetermined source separation problem, without having to assume explicitly the system to be (over-) determined, or the sources to be sparse. This is especially useful for addressing the monaural (single-channel) audio source separation problem, which is an extreme case of the underdetermined source separation problem. The task of computationally isolating acoustic sources from a mixture is extremely challenging, and recent efforts attempt to isolate speech/singing sources from monaural musical pieces or to isolate an individual's speech from a speech mixture [45] [54-59] [61] [62], and have achieved reasonable success. However, the task of separating musical sources from a monaural mixture has been, thus far, less successful in comparison.

The ability to isolate/extract individual musical components within an acoustic mixture would give an enormous amount of control over the sound. Musical pieces could be un-mixed and remixed for better musical fidelity. Signal processing, e.g. equalisation or compression, could be applied to individual instruments. Instruments could be removed from a mixture, possibly for musical students to accompany pieces of music for practice. Control over source location could be achieved in 3-D audio applications by placing the source in different locations within a 3D auditory scene.

Musical sources (instruments) have features in the frequency spectrum that are highly predictable due to the fact that they are typically constrained to specific notes (A to G# on the 12-tone musical scale) and so, frequencies are typically constrained to particular values. As such, harmonic frequencies are predictable as they can be derived from multiples of the fundamental frequency. If reliable pitch information for each source is available, harmonic frequencies for each source can be determined. With this information in hand, frequencies where harmonics from each source would overlap can be calculated. Non-overlapped harmonic frequencies in each source can therefore also be determined and non-overlapped and overlapped harmonic frequency regions in the mixture can be found, along with which particular source each non-overlapped harmonic would belong to. Existing systems [63-65] are successful in using this pitch information to identify non-overlapped harmonics and the source to which it belongs.

Polyphonic musical pieces typically have notes that complement each other (i.e. perfect 3rd, perfect 5th, minor 7th etc., explained by music theory) and so, result in a high, and regular, number of harmonics that overlap. For this reason, musical acoustic mixtures contain a larger number of overlapping harmonics in comparison to speech mixtures. Existing sound separation systems do not completely address the problem of resolving overlapping harmonics i.e. determining the contribution of each source to an overlapped harmonic. And so, because of typically higher numbers of overlapping harmonics in musical passages, musical sound separation is a difficult task and performance of existing source separation techniques has been limited. Therefore, the major challenge in musical sound separation is to effectively deal with overlapping harmonics.

A system proposed by Every and Szymanski [64] attempts to resolve overlapping harmonics by using adjacent non-overlapped harmonics to interpolate an estimate of the overlapped harmonic and so, 'fills out' the 'missing' harmonics for the spectrum of non-overlapped harmonics of each source. Nevertheless, this method relies heavily on the assumption that spectral envelopes are smooth and that amplitudes of any harmonic will have a 'middle value' of the amplitudes of the adjacent harmonics. In practice, however, spectral envelopes of real instruments rarely are smooth so this method produces varied results.

Hu [66] proposes a method of sound separation that uses onset/offset information (i.e. where performed notes start and end). Transient information in the amplitude envelope is used to determine onset/offset time by half-wave rectifying and low pass filtering the signals to obtain the amplitude envelope and the first order differential of the envelope highlights the time of sudden change in the envelope. This is a powerful cue as regions of isolated note performances can be determined. Li and Wang [63] also incorporate onset/offset information to separate sounds. However, the Li-Wang system uses the predetermined pitch information to find the onset/offset time; the time points where pitches change by at least a semi-tone are labelled appropriately as onset or offset times.

The Li-Woodruff-Wang system [67] incorporates a method utilizing common amplitude modulation (CAM) information to resolve overlapping harmonics. CAM suggests that all harmonics from a particular source have similar amplitude envelopes. The system uses the change in amplitude from the current time frame to the next of the strongest non-overlapped harmonic (in terms of a ratio), and the observed change in phase of the overlapped harmonic from the mixture to resolve the overlapped harmonic by means of least-squares estimation.

The focus of this chapter is to investigate the musical sound separation performance using pitch information and CAM principles described by Li-Woodruff-Wang [67] and proposing methods for the improvements of the system performance. The methods outlined by the pitch and CAM separation system have shown promising results, but only a small amount of research has been carried out that uses both pitch and CAM techniques together [67]. Preliminary work reveals that the pitch and CAM based system produces good results for mixtures containing long notes with considerable sustained portions e.g. a violin holding a note, but produces poor quality results for attack sections of notes, i.e. mixtures containing instruments with smaller, or no sustain sections (just attack and decay sections), e.g. a piano. Modern music typically has a high number of non-sustained note performances so the pitch and CAM method would fail with a vast number of musical pieces. In addition, the pitch and CAM method has difficulty in dealing with the overlapping harmonics, in particular, for audio sources playing similar notes.

This study aims to investigate more reliable methods of resolving harmonics for the pitch and CAM based technique of music separation which improves results, particularly for attack sections of note performances and overlapping harmonics. A method of using isolated (or relatively isolated) sections of performances in mixtures by obtaining onset/offset information is used to provide more reliable information to resolve harmonics. Such information is also used to generate a spectral template which is further used to improve the separation performance of overlapping spectral regions in the mixtures, based on the reliable information from non-overlapping regions. Implementation of the proposed methods is then attempted using a baseline pitch and CAM source separation algorithm, and system performance is evaluated.

2. Pitch and CAM system and its performance analysis

In general, the pitch and CAM system shows good performance for separating audio sources from single channel mixtures. However, according to our experimental evaluations briefly discussed below, its separation performance is limited for attack sections of notes and regions of same note performances.

We first evaluate the performance of the pitch and CAM system for separating the attack sections of music notes. To this end, we take the baseline pitch and CAM algorithm implemented in Matlab to test its performance. We use a sample database of real instrument recordings (available within ProTools music production software) to generate test files, so that the system performance on separating attack sections of notes could be evaluated. The audio file generated is a (monaural) single-channel mixture containing a melody played on a cello, and a different but complimentary melody played on a piano. The purpose of combining complimentary melodies from different sources is to generate a realistic amount of overlapping harmonics between sources, as would be found in typical musical pieces. Qualitative results show that the cello, which had long sustained portions of notes, is separated considerably well, while the attack sections of piano notes are in some cases lost as a result of the limited analysis time frame resolution. The piano has shorter notes with no sustain sections, only attacks and decays, but still contains considerable amount of harmonic content. As a result, the system performs less effectively in separating the piano source which highlights the difficulty the separation system has in isolating instruments playing short notes that are made up of regions of attack. Another

experiment on the mixture of audio sources played by clarinet and cello again confirms that the pitch and CAM system has difficulty in separating the soft attack sections of the notes played by the clarinet.

We then evaluate the system performance for the regions of same notes in the mixture. We generated a mixture containing a piano and a cello performing the same note (C4). Using the pitch and CAM system, the cello was separated from the mixture but with some artefacts and distortions. However, the system was unsuccessful in separating the piano source, and only a low level signal could be heard that did not resemble the original piano signal. In another experiment, we generated a mixture with a cello playing the note C4 and a piano playing all notes in sequence from C4 to C5 (C4, C#4, D4, D#4… etc.) and ended on the note C4. The cello was separated well from the mixture as were all notes played by the piano except the notes C4 and C5 at the both ends of the sequence. Due to the slow attack of the cello, the C4 note played by the piano at the beginning of the piece was better separated than the C4 note at the end of the sequence, as the C4 note at the beginning is more isolated. In addition, we have examined the performance of the system for mixtures with the same note and varying octaves. To this end, we generated another mixture with a cello playing the note C3 and a piano playing notes C1 to C6 in sequence and then ending on note C2. The results again show that the cello was separated well but with high distortions in sections where the piano attacks occur. The piano notes C1 and C2 were separated with some distortions but notes C3 through to C6 were almost not separated at all.

In summary, the pitch and CAM system does not perform well for recovering the sharp transients of the amplitude envelope from mixtures due to the limited time frame resolution, and it also has difficulty in separating notes with same fundamental frequencies and harmonics, caused by insufficient data for resolving the overlapping harmonics and for extracting the CAM information. For example, if one source has a pitch frequency of 50 Hz, its harmonics would occur at 100 Hz, 150 Hz, etc. If the pitch frequency of a second source is an octave higher, i.e. 100 Hz, its harmonics would occur at 200 Hz, 300 Hz, etc. As a result, the harmonics of the second source will be overlapped with those of the first source. To address these problems, we suggest two methods to improve the pitch and CAM system, respectively, isolated note and spectral template methods, which attempt to better resolve the overlapping harmonics when the information used by the pitch and CAM system is considered to be unreliable, as described next in detail.

3. Isolated note method

The proposed isolated note system, shown in Figure 1, uses note onset/offset information to determine periods of isolated performance of an instrument so that the reliable spectral information from the isolated regions can be used to resolve overlapping harmonics in the remaining note performance regions. The proposed system is based on the pitch and CAM algorithm [67], with the addition of new processing stages shown in dotted lines in Figure 1. Same to the pitch and CAM system, the inputs to the proposed system are mixture signals and pitch information supplied by a pitch tracker. The details of each block in Figure 1 are explained below.

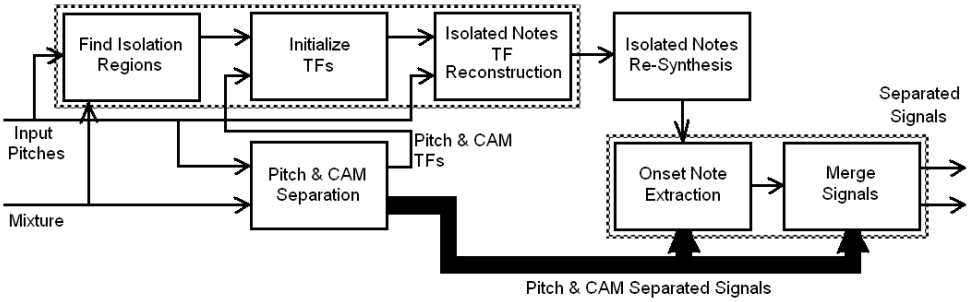


Fig. 1. Diagram of Isolated Note System.

The first processing stage is the *Pitch and CAM Separation* stage. The mixture signal is separated using the method described in [67] and by using the pitch information provided. The separated signals are used later in *Onset Note Extraction* and *Merge Signals* stages by the isolated note system. When the pitch and CAM separation is carried out the time-frequency (TF) representations of the mixture signal and the separated signals are generated which are then utilized later by the *Initialize TFs* processing stage.

The next processing stage is the *Find Isolated Regions* stage. Using input pitch information, we attempt to find time frames for each source where isolated performances of notes occur. Each time frame of each source is evaluated to determine if other sources contain pitch information (i.e. if other notes are performing during the same time frame). A list of time frames for each source is created and a flag is raised (the time frame is set to 1) if the note for the current frame and current source is isolated. Each occurrence of an isolated region (indicated by the flag) in each source is then numbered so that each region can be identified and processed independently at a later stage (achieved by simply searching through time frames and incrementing the region number at each encounter of a transition from 0 to 1 in the list of flagged time frames).

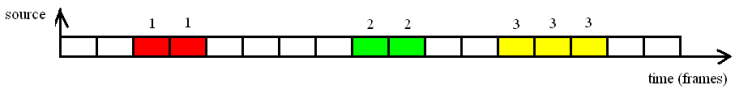
Next, we determine the non-isolated regions for the notes that contain a region of isolated note performance. For each numbered isolated region we find the corresponding non-isolated note performance and generate a new list where time frames for the non-isolated regions are numbered with the number relating to the corresponding isolated region. Note that we do not number the isolated time frames themselves in the newly generated list.

The new list is generated by searching back (previous frames) from the relevant isolated region and numbering all frames appropriately, and we then repeat by searching forward from the isolated region. Searches are terminated at endpoints of the note or at occurrences of another isolated region. Each isolated region that generates a set of corresponding non-isolated frames is saved in a new list separately, the list is then collapsed to form a final list where time frames for which we have non-isolated regions relating to two isolated regions are split halfway.

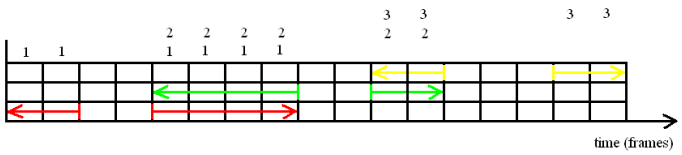
This is better illustrated by Fig. 2. Fig. 2(a) shows an occurrence of a note with three isolated regions for which information of time frames with isolated performance is determined. Fig. 2(b) illustrates that the non-isolated regions relating to each isolated

region, are found by searching forwards and backwards and terminating at endpoints of notes or an occurrence of another isolated region. Each region is stored individually. Fig. 2(c) shows the final set of regions where time frames 'belonging' to two sets are split halfway.

The TF representation of each source is formed for the isolated notes in the *Initialize TFs* stage. We initialize the TF representation by starting with an empty set of frequency information for each time frame and then by searching through the list of isolated regions. For time frames that are identified as an isolated performance of a note (from the list), we copy all frequency information for those frames directly from the mixture to the corresponding TF representation of the sources. This is shown in Fig. 3 where the time frames for the isolated performances of the note C4 (in Fig. 3(a)) are copied directly to initialize the TF representation. Fig. 3(b) shows that all of the harmonic information is copied directly from the mixture; hence all harmonics are correctly present in the initialized isolated note TF representation.



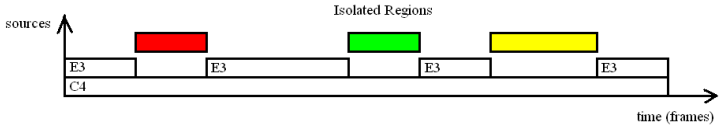
(a) Time Frames with Numbered Isolated Regions.



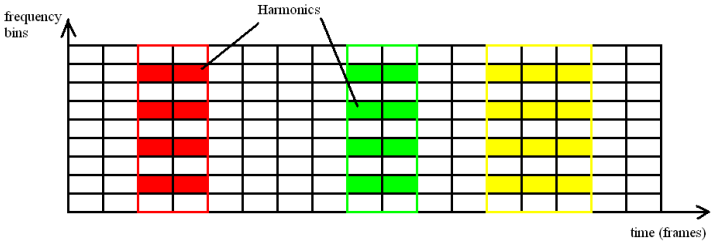
(b) Non-Isolated Regions Corresponding to Each Isolated Region.

(c) Time Frames of Non-Isolated Regions Associated with Each Isolated Region.

Fig. 2. Method Used to Determine Non-Isolated Regions of Isolated Notes.



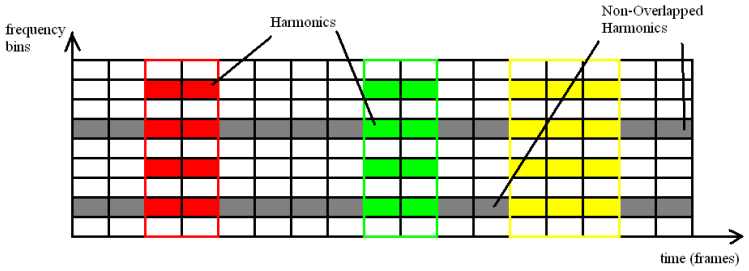
(a) Note Performance and Isolated Note Regions.



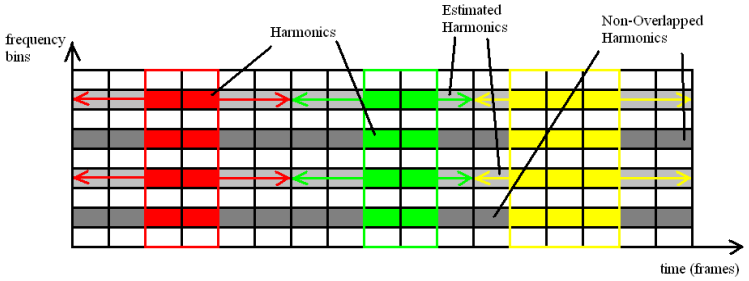
(b) Initialized TF Representations

Fig. 3. Method Used to Initialize TFs.

After the TF initialization, the *Isolated Notes TF Reconstruction* stage extends these isolated performance regions for the remaining parts of the note performances that contain isolated performance sections. Each region is evaluated in turn using information from the list of time frames for note performances which contain regions of isolated performance. The note for each time frame in the current region, and notes performed by other sources in the same time frame are determined so that a binary harmonic mask can be generated. This is then used to extract the non-overlapped harmonics for the note during sections of non-isolated performance (shown in Fig. 4(a)), which are then passed to the TF representation for relevant time frames.



(a) TFs with Non-Overlapped Harmonics Added



(b) TFs with Overlapped Harmonics Estimated Using Harmonic Information in Isolated Regions

Fig. 4. Method Used to Reconstruct TFs.

Having used non-overlapped harmonic information to update the isolated note TF representation, we can begin to estimate the overlapped harmonics for the relevant time frames. By using harmonic information available in isolated regions (for which information on all harmonics are available), amplitudes of overlapped harmonics can be estimated. Phase information for the overlapping harmonics is obtained from the corresponding harmonics in the separated TF representations found from the *Pitch and CAM Separation* stage.

As detailed earlier, each set of time frames for each source, relating to non-isolated notes containing an isolated section, are derived from time frames of corresponding isolated regions. Based on the boundary time frames, i.e. the first and last time frames of the isolated regions, we can estimate overlapped harmonic amplitudes (shown in figure 4(b)) by using the spectral information in these frames as templates. We use the first time frame frequency information in an isolated region to process previous time frames, and use the last time frame in the isolated region to process subsequent time frames. According to the CAM principle, amplitude envelopes are assumed to be the same for all harmonics. Hence, by following harmonic envelopes for the subsequent or previous time frames, we can determine the amplitude ratio $r_{t_0 \to t}$ between the template time frame t_0 and the time frame currently being processed B_t^h associated with harmonic h in time frame t

$$B_t^h = r_{t_0 \to t} B_{t_0}^h \tag{1}$$

Hence, by multiplying bins associated with an overlapped harmonic from the template frame with the ratio between frames, the amplitude for the corresponding bins in frame t can be found.

Once the TF information for notes with isolated performance regions has been constructed, it can be converted to the time domain as time-amplitude representation by the *Isolated Note Re-Synthesis* stage. The method is an adapted method used in [67]. Full frequency spectra are recreated from the half frequency spectra used in the TF representations, and the overlapped-add method is used to reconstruct the time-amplitude signals for each source. The system is designed to separate mixture signals comprising of two sources. Therefore, time domain signals of notes with isolated performance regions can be removed from the mixture signal to reveal the separated signal for the remaining source. We can simply subtract isolated note signal sample values from corresponding mixture signal sample values to generate the 'extracted' signals (performed by the *Onset Note Extraction* stage).

Finally, the *Merge Signals* stage uses isolated note signals, and the 'extracted' remaining signal, to update the separated signals obtained using the baseline pitch and CAM method. When isolated note information is available (determined by checking for a non-zero sample value) the final signal is updated with the corresponding sample in the isolated note signal for the current source being processed. The corresponding sample value is used to update the signal for the 'other' source, i.e. with the extracted signal. When isolated note information is unavailable (if a sample value of zero is encountered) the corresponding sample in the pitch and CAM separated signal, from the respective source, is used to update the final signal.

4. Spectral template method

This method aims to generate a database of spectral envelope templates of the sources from the mixtures, and then use the templates to resolve the overlapped harmonics when the pitch and CAM information is known to be unreliable. In this method, we generate a spectral envelope template for each note, using the information from the mixture. Eventually, it builds a database of spectral envelopes for all notes that are performed for each source, e.g. spectral envelopes for notes C4, E5, D# etc. The note information occurring in the mixture can be determined from the supplied pitch information. In particular, we use the non-overlapped harmonics from the most reliable sections of the mixture to fill in the spectral template for each note that appears in the mixture, where the most reliable section is regarded as the time section having the most non-overlapped harmonics for a particular instance of a note occurrence. The number of non-overlapped harmonics can vary, depending on the other notes being played simultaneously. Within this most reliable time section, the frequency spectrum at the time frame in which the largest harmonic occurs is used to train the template. Other occurrences of the note within the mixture are used to update the template for remaining unknown harmonics by analysing the ratio to adjacent non-overlapped harmonics (CAM information), based on the extraction of the 'exposed' non-overlapping harmonics. For example, when the note C5 from a source is being played together with another note G6, the 'exposed' non-overlapped harmonics of C5 can be used to train the C5 note template. Other occurrences of C5 from the same source, whilst the note A7 from the other source is being played, would 'expose' a different set of non-overlapping harmonics. These non-overlapped harmonics can be used to update the spectral template in

order to 'fill out' the unknown harmonics by using the relative amplitudes of the harmonics. This provides a 'backup' set of information for the estimation of the overlapped harmonics and also enables us to better handle situations where other information for resolving overlapping harmonics is limited or unreliable e.g. concurrent same note occurrence. Figure 5 shows the diagram of the proposed system that uses the spectral envelope model, which uses the *Pitch and CAM Separation* algorithm (developed by Li, Woodruff and Wang [67]) as a basis and adds several components shown in large red blocks (implemented in Matlab), including *Find Reliable Time Frames*, *Template Generation*, *Refine Templates*, *Update TFs*, and *Envelope Correction,* discussed next.

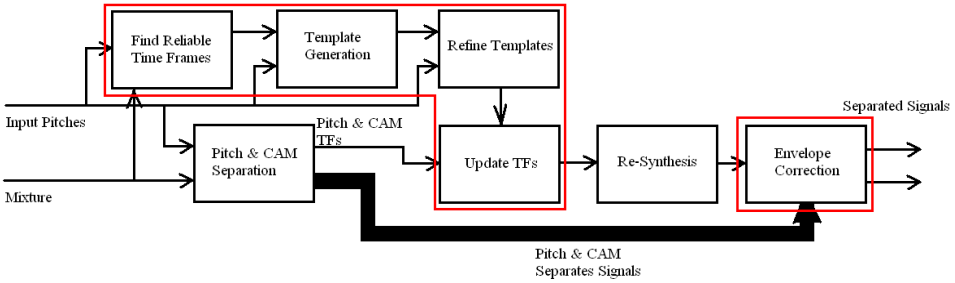


Fig.5. Diagram of the spectral template method for audio mixture separation.

The proposed spectral template system has two inputs: the mixture signal and pitch information. The input signals are the audio mixtures we attempt to separate, which can be a time-domain representation. Pitch information of each source can be extracted from the time-frequency representation of the signals, using a pitch estimator or a pitch tracker, as done in the pitch and CAM system [67]. In our proposed system, however, we use the supplied pitch information as inputs, and this essentially eliminates the influence of pitch estimation process on the separation performance. The pitch information is needed by the pitch and CAM algorithm (shown in the *Pitch and CAM Separation* stage in Figure 5) for producing an initial estimate of the sources from the TF representations of the mixtures. It is also used in *Find Reliable Time Frames* stage to determine the time frames within the TF representations that would convey the most reliable harmonic information. These time frames are then passed onto the *Template Generation* stage, and the harmonic information from these frames is used to initialize the template. In the *Refine Templates* stage, the missing harmonics of each template are estimated from the templates of other notes, when limited information is available in the mixture. The *Update TFs* stage then uses the templates at time-frames with non-overlapped harmonics to resolve the overlapped harmonics (by the *Pitch and CAM Separation* stage). These modified TF representations are passed onto the *Re-Synthesis* stage for the reconstruction of the time domain signals of each source. The *Envelope Correction* stage obtains envelope information by subtracting all but the current source from the mixture, and then use it to correct the envelope for time regions of the sources where the template was used.

In the *Pitch and CAM Separation* stage, we use the baseline algorithm developed by Li, Woodruff and Wang [67] to separate the audio mixtures, using the additional pitch contour information. More specifically, the audio mixture is transformed to the TF domain using short-time Fourier transform (STFT) with overlaps between adjacent time frames. TF

representations are generated for each separated source by the *pitch and CAM separation* algorithm, and are updated in later processing stages with improved information for time frames of unreliable information before being finally transformed back to the time domain. The separated time domain signals are also used in the *Envelope Correction* stage to obtain envelope information for the refinement of the separated signal.

In the *Find Reliable Time Frames* stage, we first find the time frames of the mixture that are most likely to yield the best set of harmonics, and we then use them to generate the spectral templates. Notes played by different instruments may have different harmonic structures, and many of them contain unreliable harmonic content. This is especially true for attack sections of many notes, due to the sharp transients and the noise content in the attack. For example, when a string is struck, its initial oscillations caused by the initial displacement will be non-periodic, and it takes a short amount of time for the string to settle into stable resonances of the instrument, and hence provide more reliable harmonic information. Some instruments may have long, slow and weak attack section, and in such a case, the harmonic content only becomes reliable at some time after the onset of the note. A similar problem also happens for notes of a short duration. In order to provide reliable frequency information for updating the note templates, we generate a list of time frames that does not include time frames containing short note performances and attack sections of note performances.

The pitch information is supplied to the *Find Reliable Time Frames* stage of the system in the form of fundamental frequencies for each source and for all time frames. The fundamental frequencies of the notes are converted into numbers representing the closest note in the 12 tone scale, i.e. C0 is 1, C#0 is 2, B0 is 12, C1 is 13 and so on up to 96 representing note B7. To find the corresponding note numbers for each frequency in the input pitch information we first determine which octave range the frequency is in by selecting an integer m such that $f_{\min} < \frac{f}{2^m} \le f_{\max}$, where the lower and upper frequency limits of the first octave (C0 to B0) are $f_{\min}$ and $f_{\max}$ respectively and f is the (fundamental) frequency value that we wish to convert to a note number. In practice, $f_{\min}$ is selected as the frequency value between C0 and the note that is one semi-tone lower (in theory, note B1), and $f_{\max}$ is selected as the frequency value between B0 and C1. The integer m can be determined by repeatedly halving the frequency until it falls within the first octave range. Once the octave range has been found, the note from A to G# on the 12-tone scale can then be found by further narrowing the searching range in terms of multiples of $f_{\min}$. In other words, we choose the integer n that satisfies the following inequality

$$\left(\sqrt[12]{2}\right)^{n-1} f_{\min} < \frac{f}{2^m} \le \left(\sqrt[12]{2}\right)^{n} f_{\min} \tag{2}$$

where m is the octave range value found previously. Once the octave range m and the note n are found, the list of note numbers at each time frame for each source can be easily calculated. From the list of notes selected, we further remove the invalid notes if they are from the attack sections of the notes, or their duration is too short. In our case, any notes whose duration is shorter than six time frames will be set to zero.

In the *Template Generation* stage, we update the spectral templates for each note with spectral information from the list of time frames that contains the valid notes obtained above. We search over each time frame in the list (also for each source in turn), and ignore the invalid time frames (with values of zero). We then determine the note performed by the current source and the notes by all other sources for the current time frame. Using such a particular note combination, we can generate binary harmonic masks to extract the non-overlapping harmonics from the TF representation of the mixtures for each of the frames. More specifically, the note performed by the current source is used to determine the frequencies of all harmonics. Notes performed simultaneously by all other sources are used to determine which of the current source harmonics are overlapped by all other sources thus, indicating the 'exposed' non-overlapping harmonics for the current note. Using such information, we can set frequency bins that are associated with non-overlapped harmonics to 1 and all other bins to 0. Firstly, the frequency of the note value for the current source must be found. According to the international standard (ISO 16 [68]), the note frequency for A4 is 440Hz, and note C0 is 57 semi-tones below A4. Hence, the frequency of C0 can be used as a basis to find the fundamental frequency of other notes such as A4 using $f = f_{C0}\left(\sqrt[12]{2}\right)^{p-1}$, where p is the note value and f_{C0} is the fundamental frequency of note C0. We then associate frequency bin b to harmonic h_i for the current source i using a similar method to that in [67], if it satisfies $\left|bf_a - h_i f\right| < \theta_1$, where θ_1 is a threshold and f_a is the frequency resolution of the TF representation (both θ_1 and f_a are determined previously in the *Pitch and CAM Separation* stage). We use a second threshold θ_2 to define the range in which the current source harmonic h_i is overlapped with any other source harmonic h_j, i.e. $\left|f_{h_j} - h_i f\right| < \theta_2$, where f_{h_j} is the frequency of harmonic h_j. Again, this is a similar method to that in [67], hence θ_2 can be chosen in the same way as used in the *Pitch and CAM Separation* stage. As a result, we can define a TF mask M_b which takes 1 if $\left|bf_a - h_i f\right| < \theta_1$, otherwise 0. This binary mask is then used to extract all non-overlapped harmonics for all time frames from the TF representation of the mixture. All the harmonic sets for the current note combination are evaluated to find the set that contains the largest amplitude harmonic, which is then used to update the note template (or simply stored if the template is empty and has not yet been initialized). We continue to go through the whole list of valid note regions, and when a new note combination is encountered, we update the note templates based on the new harmonic mask generated using the new set of 'exposed' non-overlapped harmonics. After all note combinations have been evaluated, the note templates may contain several sets of harmonics for each note combination. If this happens, we merge them to create a final set for the note template. Note that one may wish to apply scaling to each set of harmonic templates to ensure harmonics are of correct magnitude when merging the template.

As done in the *Refine Templates* stage, the spectral templates generated, are further refined and improved by using information from all the templates. The reason that the spectral templates need to be refined is because for some notes, there may be only a limited set of non-overlapped harmonics, as some harmonics may not be available in the mixture. To improve the templates, harmonic information from other note templates that are available within a specified range of notes is used. Spectra of other note templates are pitch shifted to

match the note we intend to improve, so that information for correlating harmonics can be obtained (after harmonics are aligned). However, spectral quality tends to deteriorate as the degree of pitch shifting increases. Therefore we first use the templates of notes that are closest in frequency to the note template for which we wish to improve, and then continue with templates of decreasing quality. In addition, lower frequency note templates yield higher quality spectra when the pitch is shifted up to match the frequency of the note template we wish to improve, and vice versa. Hence, we limit the range of notes and also the number of note templates to be used for improving note templates. This essentially excludes note templates that have been excessively pitch shifted, and also improves computational efficiency of the proposed system.

In the *Update TFs* stage, we update the TF representations of the separated sources from the *Pitch and CAM Separation* stage, using the note templates. Pitch information is used to determine, for each source, the time frames where reliable non-overlapped harmonics are unavailable for separation. As already mentioned, if a source is playing a note which is one octave lower than a note played by another source, the former one would have every other harmonic overlapped whereas the harmonics of the latter one would be totally overlapped by those of the former one. As a consequence, no reliable information is available to resolve the overlapped harmonics of the latter source. However, there are many other note combinations, leading to unavailable non-overlapping harmonics to be used to resolve the overlapped harmonics, e.g. when one source performs a note 7 semi-tones higher (perfect fifth interval) than the other it would result in every third harmonic of the latter source being overlapped by the former one. Of course, it would be exhaustive to find all possible combinations of notes that result in all of the source harmonics being overlapped. Using pitch information is an efficient way to calculate the resulting number of overlapped harmonics at each time frame for each source. The number of overlapping harmonics $\varphi_i(t)$ for source i at time frame t can be determined by finding the number of harmonics in a complete set $H_{N_i}(t)$ that is not in the set of non-overlapped harmonics $\tilde{H}_{N_i}(t)$ based on the pitch information of note $N_i(t)$.

We use the same method discussed above to generate binary masks, using current note information and information on all other notes that are performed simultaneously. We also create a binary mask with a complete set of harmonics from which the mask with non-overlapped harmonics is subtracted. This gives a mask containing harmonics that are overlapped. Evaluating the magnitude at bins closest to the expected harmonic frequencies allows the number of overlapped harmonics present to be determined. For all t where $\varphi_i(t) = 0$ i.e. time frames for source i that have no reliable information for source separation, frequency spectra for the respective note templates are used to replace the frequency spectra in the TF representation of the separated source.

The *Re-Synthesis* stage, adapted from [67], involves the reconstruction of the time domain signals from the TF representations for each source. Specifically, symmetric frequency spectra are created from the half spectra used in the TF representations and the overlap-add method is used to generate the time domain signals.

No amplitude envelope information has been conveyed in the note templates for refining the separated sources. Hence, in the *Envelope Correction* stage, for the time regions with

unresolved overlapped harmonics, the amplitude envelopes of the separated sources will be corrected. All sources that have been separated (in the *Pitch and CAM Separation* stage) except the current source, for which the envelope is being corrected, are removed from the original mixture signal. The remaining signal would then be a crude representation of the source we are attempting to correct as most of the high energy components from all other sources are removed. The envelope of the remaining signal is found by finding peaks of absolute amplitude values. We detect peaks at time instances where the first order derivative of the absolute time-amplitude signal is zero. The envelopes of the separated sources are then adjusted by applying a certain amount of scaling determined by the desired envelope obtained above.

5. System evaluation

5.1 Evaluation method

The system is evaluated using test signals specifically designed to highlight differences between the proposed systems and the original pitch and CAM separation system. The proposed systems aim to address the weak points of the pitch and CAM system, i.e. the lack of time domain detail arising from poor separation of attack regions of notes, and its difficulty in resolving the overlapping harmonics due to similar note performances. Hence, tests were designed to evaluate differences in these particular points between the proposed systems and the original system, rather than an evaluation of overall performance of the system.

For the proposed isolated note system, test signals which were generated using real instrument recordings with different musical scores, contain isolated performances of notes in order to show the effectiveness of the proposed system. The isolated note system aims to better resolve attack sections of notes for which the pitch and CAM system performs poorly. Hence, instruments with fast attacks and relatively higher energy in the higher frequency range (of the attacks), e.g. instruments that are struck, or particular instruments that are plucked were selected for the test signals. Two test signals meeting these criteria were generated; the first signal (test signal 1) was a two-source mixture containing a cello and a piano performance, the cello was played throughout the signal and the piano had sections of performance interspersed with sections of silence giving the cello regions of isolated performance. The second signal (test signal 2) was also a two-source mixture containing a string section and a guitar performance, again, the string section was played throughout the test signal and the guitar had interspersed sections of silence. Both test mixtures were created by mixing clean source signals (16-bit, 44100Hz sample rate).

For the spectral template system, two test signals with the same musical score are generated containing sections with the same note performance and also sections with sufficient information to train the templates. The first piece was a two source mixture of a cello and a piano, the second piece was a two source mixture of a cello and a clarinet (both pieces approximately four seconds long at 16 bit, 44100 kHz sampling rate). All the test signals were created using ProTools music production software and instruments were selected to avoid synthesized replications to achieve performances as realistic as possible (this avoids signals being created with stable frequency spectra for note performances). A database of real recordings of instruments within the music production software was used to generate the test signals. Pitch and CAM separation was performed with default values.

System performance is evaluated by calculating the SNR for the pitch and CAM system and the proposed system with each test signal using

$$SNR(dB) = 10\log_{10}\frac{\sum_{n}\left(x[n]\right)^2}{\sum_{n}\left(x[n]-\hat{x}[n]\right)^2} \tag{3}$$

where $x[n]$ is the original signal and $\hat{x}[n]$ is the separated signal (n is the sample index). This allows us to quantify the sample-wise resemblance between the clean source signals and the separated signals generated by each of the systems.

For the evaluation of the isolated system, a direct comparison of SNR values for both systems would reveal the gains made by the isolated note system. However, differences in the attack sections only are difficult to quantify when evaluating the entire signal as they make up only a small proportion of the test signal. Hence, we expect the differences in perceptual quality to be more significant (i.e. differences would be heard, but are not represented as well in comparison using SNR measurements). Therefore, a listening test was also performed to observe the perceptual difference between the separated signals obtained using the pitch and CAM and the isolated note methods. Test signals for the listening test were generated by including the original clean source signal, followed by a one second silence, and then the separated signal allowing for a direct comparison to be made between the clean source and separated signals. 26 participants were asked to score the signals from 0 to 5, with 0 being extremely poor and 5 being perceptually transparent (with reference to the original signal). Scores were based on the details of attack sections as well as overall separation performance between the two systems (i.e. which system 'sounds better') all test signals were presented in a random order for each participant.

For the evaluation of the spectral template system, the separated signals are modified to remove the pitch and CAM sections so that the signals contain only same note performances, and the influence of the pitch and CAM results is ignored. Test signals are created by including the original signal at the start, followed by a one second silence, and then followed by the separated signal; this allows the listener to hear the original before hearing the separated signal so a direct comparison can be made. Test signals were generated for both pitch and CAM and note template systems. All test signals were played in a random order so that identification of each system remains unknown and cannot be anticipated. Signals were allowed to be repeated as many times as needed to assess signal quality.

5.2 Results

The results of the isolated note system are shown in Tables 1 and 2. When comparing results for test signal 1, source 1 (cello), we observe a reduction of -3.75 dB in SNR between the two systems. Nevertheless, this source contains sections of isolated performance which we use to better separate attack sections of source 2 (for which this study concerns). As can be seen for source 2 (piano), SNR of the proposed system is 15.08 dB higher than the pitch and CAM system, so a significant gain in separation performance is achieved. Looking at SNR results for test signal 2 source 1 (string section) we see a marginal increase of 0.34 dB in separation performance from the isolated note system, again, this source contains the isolated region of

performance which is used to improve separation of source 2. For source 2 (guitar), we see a significant improvement in separation performance by the isolated note system with a SNR 8.44 dB higher than the pitch and CAM system.

Test Signal	Source	Pitch and CAM System	Isolated Note System
1	1	19.04	15.29
	2	5.87	20.95
2	1	15.78	16.09
	2	3.63	12.07

Table 1. SNR (dB) results for Isolated Note System as compared with the pitch and CAM system.

Test Signal	Source	Pitch and CAM Mean Score	Isolated Note Mean Score
1	1	4.88	4.73
	2	2.50	4.50
2	1	3.85	3.69
	2	1.65	3.54

Table 2. Listening test results for Isolated Note System as compared with the pitch and CAM system.

For test signal 1 we can see similar mean opinion scores for separation of source 1 by both systems suggesting a similar level of separation performance between the two systems. However, listening test results suggest a significant improvement of separation performance by the isolated note system for source 2. For test signal 1 and source 2, the pitch and CAM system achieved a mean score of 2.50 and the isolated note system achieved a mean score of 4.50. Again, the isolated note system achieved similar separation performance compared to the pitch and CAM system for test signal 2, source 1, while giving a significant improvement for source 2. The pitch and CAM system achieved a mean score of 1.65 whereas the isolated note achieved a higher mean score of 3.54. Both SNR and listening test results indicate that the note isolation separation system achieves better separation performance. We can see significant quantitative gains from the SNR results for signals with fast attacks (source 2 in both test signals 1 and 2). Qualitative results from the listening test also show significant perceptual gains obtained in the separation of attack sections in addition to overall separation.

The results of the spectral template system are summarised in Tables 3 and 4. Table 3 shows SNR results for the proposed note template separation system compared with the pitch and CAM separation system. We can see that for both test signals, we have the same separation performance for source 1 (cello). Sufficient harmonic information is available for source 1 to resolve overlapping harmonics so the note template system also uses the pitch and CAM method to separate the signal which is why the same performance result can be observed. However, for source 2 (piano), SNR results appear to be poor. For test signal 1 we see that the pitch and CAM system has a SNR of 0.79 dB whereas the note template system has a SNR of -2.35 dB, suggesting that the level of noise introduced by the system is greater than the level of input signal. Likewise, test signal 2 shows poor SNR results for source 2, the pitch and CAM system has a SNR of 2.90 dB while the note template system has a SNR of -3.65 dB.

Test Signal	Source	Pitch and CAM System	Note Template System
1	1	2.62	2.62
	2	0.79	-2.35
2	1	7.79	7.79
	2	2.90	-3.65

Table 3. SNR (dB) results for Note Template System as compared with the pitch and CAM system.

Test Signal	Source	Pitch and CAM System	Note Template System
1	1	4.08	3.77
	2	1.96	0.92
2	1	4.77	4.81
	2	1.65	0.92

Table 4. Listening test results for Note Template System as compared with the pitch and CAM system.

Table 4 shows average results for the listening test for the pitch and CAM separation system and the note template separation system. For test signal 1, source 1, we see a mean score of 4.08 for the pitch and CAM separation system and a mean score of 3.77 for the note template system despite the same pitch and CAM separated signal being used by both systems, as explained earlier. For test signal 1, source 2, we see a mean score of 4.77 for the pitch and CAM system. We see a reduction of the score for the note template system, with a mean score of 0.92. Comparing scores for test signal 2, similar scores for source 1 can be seen for both systems, with the pitch and CAM system scoring a mean of 4.77 and the note template system scoring a mean of 4.81. Again, both systems use the pitch and CAM separated signals for source 1, as explained earlier. The score for the note template system is lower than the score for the pitch and CAM system, for test signal 2, source 2; we see a mean score of 1.65 for the pitch and CAM system and a mean score of 0.92 for the note template system. The spectral template system does not work as promising as we would have expected, due to the following possible reasons. The templates trained from mixtures may not be accurate enough to represent the sources, because of the limited number of non-overlapped harmonics and isolated notes within the mixture. Using clean music source data (instead of the monaural mixture) to train the templates may be able to mitigate this problem and further to improve the results. Also, in the proposed template systems, pitch shifting which was used to fill up the missing notes that are not available in the mixture, apparently introduces errors in harmonic estimation. These are interesting points for future investigation.

6. Conclusions

We have presented two new methods for music source separation from monaural mixture using the isolated note information and note spectral template, both evaluated from the sound mixture. The proposed methods were designed to improve the separation performance of the baseline pitch and CAM system especially for the separation of attack sections of notes, and overlapping time-frequency regions. In the

pitch and CAM system, the fast attack sections are almost completely lost in the separated signals, resulting in poor separation results for the transient part of the signal. In the proposed isolate note system, accurate harmonic information available in the isolated regions is used to reconstruct harmonic content for the entire note performance, and so, the harmonic content can be removed from the mixture to reveal the remaining note performance (in a two-source case). The isolated note system has been shown to be successful in improving the separation performance of attack sections of notes, offering a large improvement in separation quality over the baseline system. In the proposed note template system, the overlapping time-frequency regions of the mixtures are resolved using the reliable information from the non-overlapping regions of the sources, based on the spectral template matching. Preliminary results show that the spectral templates evaluated from the mixtures can be noisy and may degrade the results. Using spectral template generated directly from clean training data (i.e. containing single signals, instead of mixtures) has the potential to improve the system performance which will be our future study.

7. Future directions

We have studied the potentials of using spectral template and isolated note information for music sound separation. A major challenge is however to identify the regions from which the note information can be regarded as reliable and thereby used to estimate the note information for the unreliable and overlapped regions. Under noisy and multiple source conditions, more ambiguous regions may be identified, and using such information may further distort the separation results. Pitch information is relatively reliable under noisy conditions and can be used to improve the system performance [81]. Another potential direction is to use the property of the sources and noise/interferences, such as sparseness, to facilitate the identification of the reliable regions within the mixture that can be used to estimate the sources [74-77]. This is mainly due to the following three reasons. Firstly, as mentioned earlier, music audio can be made sparser if it is transformed into another domain, such as the TF domain, using an analytically pre-defined dictionary such as discrete Fourier transform (DFT) or discrete cosine transform (DCT) [69] [70]. Recent studies show that signal dictionaries directly adapted from training data using machine learning techniques, based on some optimisation criterion (such as the reconstruction error regularised by a sparsity constraint), can offer better performance than the pre-defined dictionary [71] [72]. Secondly, the sparse techniques using learned dictionary have been shown to possess certain denoising capability for corrupted signals [72]. Thirdly, identification of reliable regions from sound mixtures, and estimation of the probability of each TF point dominated by a source can be potentially cast as an audio-inpainting [73] or matrix completion problem. This naturally links the two important areas: source separation and sparse coding. Hence, the emerging algorithms developed in the sparse coding area could be potentially used for the CASA based monaural separation system. Separating music sources from mixtures with uncertainties [78] [79], such as under the condition of unknown number of sources, is also a promising direction for future research, as required in many practical applications. In addition, online optimisation will be necessary when the separation algorithms operate on resource limited platforms [80].

8. References

[1] Jutten, C., & Herault, J. (1991). Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture, *Signal Processing*, vol. 24, pp. 1-10.

[2] Cardoso, J.-F., & Souloumiac, A. (1993). Blind Beamforming for Non Gaussian Signals, *IEE Proc. F, Radar Signal Processing*, vol. 140, no. 6, pp. 362-370.

[3] Comon, P. (1994). Independent Component Analysis: a New Concept?, *Signal Processing*, vol. 36, no. 3, pp. 287–314.

[4] Bell, A. J., & Sejnowski, T. J. (1995). An Information Maximization Approach to Blind Separation and Blind Deconvolution, *Neural Computation*, vol. 7, no. 6, pp. 1129-1159.

[5] Amari, S.-I., Cichocki, A., & Yang, H. (1996). A New Learning Algorithm for Blind Signal Separation, *Advances Neural Information Processing System*, vol. 8, pp. 757–763.

[6] Cardoso, J.-F. (1998). Blind Signal Separation: Statistical Principles, *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025.

[7] Lee, T.-W. (1998). *Independent Component Analysis: Theory and Applications*. Boston, MA: Kluwer Academic.

[8] Haykin, S. (2000). *Unsupervised Adaptive Filtering*, Volume 1, Blind Source Separation. New York: Wiley.

[9] Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis*. New York: Wiley.

[10] Cichocki, A., & Amari, S.-I. (2002). *Adaptive Blind Signal and Image Processing, Learning Algorithm and Applications*. New York: Wiley.

[11] Cardoso, J. F., & Laheld, B. (1996). Equivariant Adaptive Source Separation, *IEEE Transactions Signal Processing*, vol. 44, no. 12, pp. 3017–3030.

[12] Belouchrani, A., Abed-Meraim, K., Cardoso, J., & Moulines, E. (1997). A Blind Source Separation Technique Using Second-Order Statistics, *IEEE Transactions Signal Processing*, vol. 45, no. 2, pp. 434–444.

[13] Thi, H., & Jutten, C. (1995). Blind Source Separation for Convolutive Mixtures, *Signal Processing*, vol. 45, pp. 209-229.

[14] Smaragdis, P. (1998). Blind Separation of Convolved Mixtures in the Frequency Domain, *Neurocomputing*, vol. 22, pp. 21–34.

[15] Parra, L., & Spence, C. (2000). Convolutive Blind Source Separation of Nonstationary Sources, *IEEE Transactions Speech Audio Processing*, vol. 8, no. 3, pp. 320–327.

[16] Rahbar, K., & Reilly, J. (2001). Blind Source Separation of Convolved Sources by Joint Approximate Diagonalization of Cross-Spectral Density Matrices, in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing*, Utah, USA.

[17] Davies, M. (2002). Audio Source Separation, in *Mathematics in Signal Separation V*. Oxford, U.K.: Oxford Univ. Press.

[18] Sawada, H., Mukai, R., Araki, S., & Makino, S. (2004). A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation, *IEEE Transactions Speech and Audio Processing*, vol.12, no. 5, pp. 530-538.

[19] Wang, W., Sanei, S., & Chambers, J.A. (2005). Penalty Function Based Joint Diagonalization Approach for Convolutive Blind Separation of Nonstationary Sources, *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1654-1669.

[20] Sawada, H., Araki, S., & Makino, S. (2010). Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment, *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, pp. 516-527.

[21] Pedersen, M., Larsen, J., Kjems, U., & Parra, L. (2007). A Survey on Convolutive Blind Source Separation Methods, *Handbook on Speech Processing and Speech Communication*, Springer.

[22] Belouchrani, A., & Amin, M. G. (1998). Blind Source Separation Based on Time-Frequency Signal Representations, *IEEE Transactions Signal Processing*, vol. 46, no. 11, pp. 2888–2897, 1998.

[23] Chen, S., Donoho, D. L., & Saunders, M. A. (1998). Atomic Decomposition by Basis Pursuit, *SIAM Journal Scientific Computing*, vol. 20, no. 1, pp. 33-61.

[24] Lee, T., Lewicki, M., Girolami, M., & Sejnowski, T. (1998), Blind Source Separation of More Sources Than Mixtures Using Overcomplete Representations, *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87-90.

[25] Lewicki, M. S., & Sejnowski, T. J. (1998). Learning Overcomplete Representations, *Neural Computation*, vol. 12, no. 2, pp. 337–365.

[26] Bofill, P., & Zibulevsky, M. (2001). Underdetermined Blind Source Separation Using Sparse Representation, *Signal Processing*, vol. 81, pp. 2253–2362.

[27] Zibulevsky, M., & Pearlmutter, B. A. (2001). Blind Source Separation by Sparse Decomposition in a Signal Dictionary, *Neural Computation*, vol. 13, no. 4, pp. 863–882.

[28] Li, Y., Amari, S., Cichocki, A., Ho, D. W. C., & Xie, S. (2006). Underdetermined Blind Source Separation Based on Sparse Representation. *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 423-437.

[29] He, Z., Cichocki, A., Li, Y., Xie, S., & Sanei, S. (2009). K-Hyperline Clustering Learning for Sparse Component Analysis. *Signal Processing*, vol. 89, no. 6, pp. 1011-1022.

[30] Yilmaz, O., & Richard, S. (2004). Blind Separation of Speech Mixtures via Time-Frequency Masking, *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830-1847.

[31] Wang, D.L. (2005). On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis. In Divenyi P. (ed.), *Speech Separation by Humans and Machines*, pp. 181-197, Kluwer Academic, Norwell MA.

[32] Mandel, M. I., Weiss, R. J., & Ellis, D. P. W. (2010). Model-based Expectation Maximisation Source Separation and Localisation, *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, pp. 382-394.

[33] Duong, N. Q. K., Vicent, E., & Gribonval, R. (2010). Under-determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model, *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, pp. 1830-1840.

[34] Plumbley, M. D. (2003). Algorithms for Nonnegative Independent Component Analysis, *IEEE Transactions on Neural Networks*, vol. 14, no. 2, pp. 534–543.

[35] Kim, T., Attias, H., & Lee, T.-W. (2007). Blind Source Separation Exploiting Higher-Order Frequency Dependencies, *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, pp. 70-79.

[36] Comon, P., & Jutten, C., eds. (2010). *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press.

[37] Smaragdis, P. (2004). Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. In *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation*, Grenada, Spain.

[38] Schmidt, M. N., & Mørup, M. (2006). Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation, in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation*, Charleston, USA.

[39] Wang, W., Cichocki, A., & Chambers, J. A. (2009). A Multiplicative Algorithm for Convolutive Non-negative Matrix Factorization Based on Squared Euclidean Distance, *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2858-2864.

[40] Ozerov, A., & Févotte, C. (2010). Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation, *IEEE Transactions on Audio, Speech and Language Processing*.

[41] Mysore, G., Smaragdis, P., & Raj, B. (2010). Non-Negative Hidden Markov Modeling of Audio with Application to Source Separation. In *Proceedings of the 9th international conference on Latent Variable Analysis and Signal Separation (LCA/ICA)*. St. Malo, France.

[42] Ozerov, A., Févotte, C., Blouet, R., & Durrieu, J.L. (2011). Multichannel Nonnegative Tensor Factorization with Structured Constraints for User-guided Audio Source Separation, *Proceedings of the IEEE International Conference Acoustics, Speech and Signal Processing*, Prague, Czech Republic.

[43] Wang, W. & Mustafa, H. (2011). Single Channel Music Sound Separation Based on Spectrogram Decomposition and Note Classification, in *Computer Music Modelling and Retrieval*, Springer.

[44] Cichocki, A., Zdunek, R., Phan, A.H., & Amari, S. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley.

[45] Brown, G.J., & Cooke, M.P. (1994). Computational Auditory Scene Analysis, *Computer Speech and Language*, vol. 8, pp. 297–336.

[46] Wrigley, S. N., Brown, G. J., Renals, S., & Wan, V. (2005). Speech and Crosstalk Detection in Multi-Channel Audio, *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84-91.

[47] Palomäki, K. J., Brown, G. J., & Wang, D. L. (2004). A Binaural Processor for Missing Data Speech Recognition in the Presence of Noise and Small-Room Reverberation, *Speech Communication*, vol. 43, no. 4, pp. 361-378.

[48] Wang, D.L., & Brown, G. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE.

[49] Shao, Y., & Wang, D.L. (2009). Sequential Organization of Speech in Computational Auditory Scene Analysis. *Speech Communication*, vol. 51, pp. 657-667.

[50] Hu, K., & Wang, D.L. (2011). Unvoiced Speech Segregation from Nonspeech Interference via CASA and Spectral Subtraction. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1600-1609.

[51] Xu, T., & Wang, W. (2009). A Compressed Sensing Approach for Underdetermined Blind Audio Source Separation with Sparse Representations, in *Proceedings of the IEEE International Workshop on Statistical Signal Processing*, Cardiff, UK.

[52] Xu, T., & Wang, W. (2010). A Block-based Compressed Sensing Method for Underdetermined Blind Speech Separation Incorporating Binary Mask, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, USA.

[53] Xu, T., & Wang, W. (2011). Methods for Learning Adaptive Dictionary for Underdetermined Speech Separation, in *Proceedings of the IEEE 21st International Workshop on Machine Learning for Signal Processing*, Beijing, China.

[54] Kim, M., & Choi, S. (2006). Monaural Music Source Separation: Nonnegativity, Sparseness, and Shift-Invariance, in *Proceedings of the IEEE International Conference on Independent Component Analysis and Blind Signal Separation*, Charleston, USA.

[55] Virtanen, T. (2006). *Sound Source Separation in Monaural Music Signals*, PhD Thesis, Tampere University of Technology.

[56] Virtanen, T. (2007). Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066-1074.

[57] Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hawaii, USA.

[58] Richard, G., & David, B. (2009). An Iterative Approach to Monaural Musical Mixture De-Soloing, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan.

[59] Klapuri, A., Virtanen, T., & Heittola, T. (2010). Sound Source Separation in Monaural Music Signals Using Excitation-Filter Model and EM Algorithm, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, USA.

[60] Bregman, A. S. (1990). *Auditory Scene Analysis*, MIT Press.

[61] Li, Y. & Wang, D. L. (2007). Separation of Singing Voice From Music Accompaniment for Monaural Recordings, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1475-1487.

[62] Parsons, T. W. (1976). Separation of Speech from Interfering Speech By Means of Harmonic Selection, *Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911-918, 1976.

[63] Li, Y. & Wang, D. L. (2009). Musical Sound Separation Based on Binary Time-Frequency Masking, *EURASIP Journal on Audio, Speech and Music Processing*, article ID 130567.

[64] Every, M. R. & Szymanski, J. E. (2006). Separation of Synchronous Pitched Notes by Spectral Filtering of Harmonics, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1845-1856.

[65] Virtanen, T. & Klapuri, A. (2001). Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation, in *Proceedings of the IEEE Workshop on Applications of Signal Processing in Audio and Acoustics*, pp. 83-86.

[66] Hu, G. (2006). *Monaural Speech Organization and Segregation*, Ph.D. Thesis, The Ohio State University, USA.

[67] Li, Y., Woodruff, J. & Wang, D. L. (2009). Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1361-1371.

[68] ISO. *Acoustics – Standard Tuning Frequency (Standard Musical Pitch)*, ISO 16:1975, International Organization for Standardization, Geneva, 1975.

[69] Nesbit, A., Jafari, M. G., Vincent, E., & Plumbley, M. D. (2010). Audio Source Separation Using Sparse Representations. In W. Wang (Ed), *Machine Audition: Principles, Algorithms and Systems*. Chapter 10, pp.246-264. IGI Global.

[70] Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R., & Davies, M. E. (2010). Sparse Representations in Audio and Music: from Coding to Source Separation, *Proceedings of the IEEE*, vol. 98, pp. 995-1005.

[71] Dai, W., Xu, T. & Wang, W. (2012). Dictionary Learning and Update based on Simultaneous Codeword Optimisation (SIMCO), *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan.

[72] Dai, W., Xu, T., & Wang, W. (2011). Simultaneous Codeword Optimisation (SimCO) for Dictionary Update and Learning, *arXiv:1109.5302*.

[73] Adler, A., Emiya V., Jafari, M.G., Elad, M., Gribonval, G., & Plumbley, M.D. (2012). Audio Inpainting, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 922-932.

[74] Wang, W. (2011). *Machine Audition: Principles, Algorithms and Systems*, IGI Global Press.

[75] Jan, T. & Wang, W. (2011). Cocktail Party Problem: Source Separation Issues and Computational Methods, in W. Wang (ed), *Machine Audition: Principles, Algorithms and Systems*, IGI Global Press, pp. 61-79.

[76] Jan, T., Wang, W., & Wang, D.L. (2011). A Multistage Approach to Blind Separation of Convolutive Speech Mixtures. *Speech Communication*, vol. 53, pp. 524-539.

[77] Luo, Y., Wang, W., Chambers, J. A., Lambotharan, S., & Prouder, I. (2006). Exploitation of Source Non-stationarity for Underdetermined Blind Source Separation With Advanced Clustering Techniques, *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2198-2212.

[78] Adiloglu, K. & Vincent, E. (2011). An Uncertainty Estimation Approach for the Extraction of Source Features in Multisource Recordings, in *Proceedings of the European Signal Processing Conference*, Barcelona, Spain.

[79] Adiloglu, K., & Vincent, E. (2012). A General Variational Bayesian Framework for Robust Feature Extraction in Multisource Recordings, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan.

[80] Simon, L. S. R., & Vincent, E. (2012). A General Framework for Online Audio Source Separation, in *Proceedings of the International conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel.

[81] Hsu, C.-L., Wang, D.L., Jang J.-S.R., & Hu, K. (2012). A Tandem Algorithm for Singing Pitch Extraction and Voice Separation from Music Accompaniment, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1482-1491.