RECONSTRUCTING COMPLETELY OVERLAPPED NOTES FROM MUSICAL MIXTURES

Jinyu Han *Bryan Pardo*

Northwestern University
2133 Sheridan Road, Evanston, IL 60208, USA.

ABSTRACT

In mixtures of musical sounds, the problem of overlapped harmonics poses a significant challenge to source separation. *Common Amplitude Modulation (CAM)* is one of the most effective methods to resolve this problem. It, however, relies on non-overlapped harmonics from the same note being available. We propose an alternate technique for harmonic envelope estimation, based on *Harmonic Temporal Envelope Similarity (HTES)*. We learn a harmonic envelope model for each instrument from the non-overlapped harmonics of notes of the same instrument, wherever they occur in the recording. This model is used to reconstruct the harmonic envelopes for overlapped harmonics. This allows reconstruction of completely overlapped notes. Experiments show our algorithm performs better than an existing system based on *CAM* when the harmonics of pitched instruments are strongly overlapped.

Index Terms— Music Source Separation, Common Amplitude Modulation, Harmonic Temporal Envelope Similarity

1. INTRODUCTION

Musical source separation is the process of isolating individual parts from audio containing multiple concurrent musical instruments. A solution to this problem has potential applications to many tasks, such as music transcription, content-based analysis, and query by humming. Almost all music source separation systems have to deal with harmonics that overlap in the time-frequency domain. Overlapped harmonics are very common in tonal music based on the 12-tone equal tempered scale. This includes most Western classical, jazz, pop, folk, blues and rock music. Resolving this problem is key to music source separation.

Systems inspired by *computational auditory scene analysis* (CASA) [1] rely on various assumptions about the audio to deal with overlapped harmonics. *Spectral smoothness* [2, 3] assumes that the spectral envelope of every instrument sound is smooth. The amplitude of an overlapped harmonic is estimated from the amplitudes of the neighboring non-overlapped harmonics using various weighting techniques. *Common Amplitude Modulation* (CAM) [4, 5] assumes that the amplitude envelopes of different harmonics of the same note tend to be similar. Assuming CAM, the amplitude envelope of the overlapped harmonic is approximated from the envelopes of the non-overlapped harmonics of the same note.

Tonal music makes extensive use of simultaneous instruments, playing consonant intervals, such as the octave (F0 of instrument 1 is twice F0 of instrument 2) or the fifth (F0 of instrument 1 is 3/2 F0 of instrument 2). The result is that the harmonics of the lower pitched instrument partially or completely (in the case of octaves) overlap the harmonics of the higher pitched instrument. The above-mentioned methods based on CAM and spectral smoothness fail to deal with the complete overlap problem.

We proposed a new separation system, illustrated in Fig. 1, to solve the overlap problem. We address the separation problem in four stages. The first stage is *Harmonic Mask Estimation* where pitches are used to construct harmonic masks to identify the overlapped harmonics. In the second stage, *Harmonic Envelope Estimation*, the harmonic envelopes of the overlapped harmonics are estimated. The amplitudes and phases of overlapped harmonics are estimated in the stage of *Harmonic Phase and Amplitude Estimation*. The resulting spectrogram estimate for each source is converted to a time-domain estimate by overlap-add in the *Re-synthesis* stage.

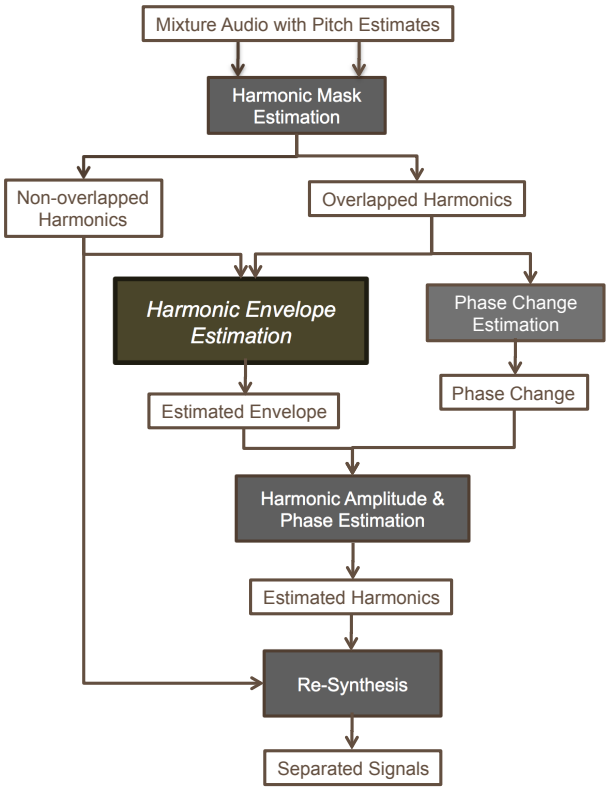

Fig. 1. System overview

We propose a new method to solve the "complete overlap" problem. It is inspired by scene completion [6] that patches up holes

This work was supported in part by National Science Foundation award 0643752.

in image using sections of other images. We utilize the property that notes played by the same instrument within a short period of time (e.g. within musical phrase boundaries) have similar harmonic envelopes. In a mixture containing several instruments, we learn the harmonic envelope model for each instrument from the non-overlapped harmonics of notes played by that source throughout the recording. To recover 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 *CAM* but also deals with completely overlapped notes. Scene completion is done based on the audio within the file to be processed. There is no need for an outside library of similar sounds. We focus our paper on the stage *Harmonic Envelope Estimation* due to the length constraint. A full description of our system is in [7].

2. HARMONIC TEMPORAL ENVELOPE SIMILARITY

Existing work based on *Common Amplitude Modulation* assumes the harmonic amplitude envelopes from the same note are correlated. *CAM* holds most of the time for the first few strong harmonics and fails to hold for those with weak energy. Fig. 2 showed the amplitude envelopes of first 20 harmonics of a clarinet. We can see the envelopes of the first 8 harmonics that contain 96% of the energy, do share the same general modulation trend, while the envelopes of the remaining harmonics have little correlation with each other.

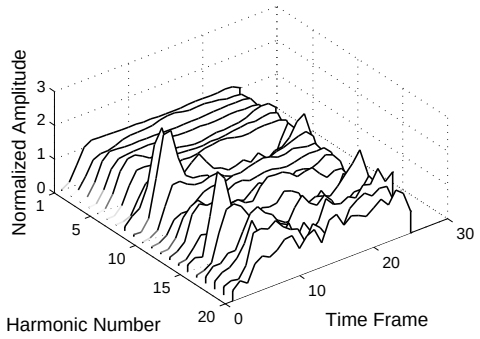

Fig. 2. First 20 harmonic envelopes of the note F4 note played by a clarinet. Amplitude envelopes are normalized in volume so that the shapes are easier to see.

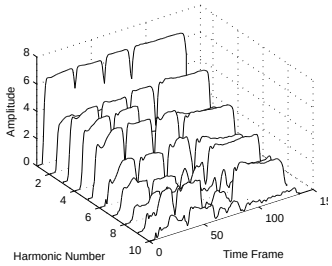

Fig. 3. The first 10 harmonics of four consecutive notes of 400Hz, 375Hz, 300Hz and 330Hz played by a clarinet

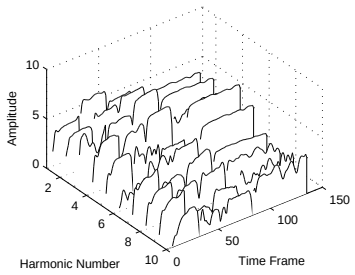

Fig. 4. The first 10 harmonics of four consecutive notes of 132Hz, 147Hz, 197Hz and 100Hz played by a bassoon

The harmonic envelope of notes played by different instruments may be very different, but the envelopes of different notes played by the same instrument within a short period of time usually show great resemblances. Fig. 3 and 4 showed the first 10 harmonic envelopes of four consecutive notes played by a clarinet and bassoon. Although these notes have different pitches and lengths, their amplitude envelopes of the strong harmonics evolve similarly across notes. We call the similarity of harmonic envelope among different notes of the same instrument *Harmonic Temporal Envelope Similarity (HTES)*.

Fig.5 shows the amplitude envelope of the first harmonic of nine notes played on a clarinet as a solid line. Periods where the first harmonic was strongly overlapped by another instrument are indicated by asterisks along the top horizontal axis. Estimation of the harmonic envelope by CAM and by HTES are plotted as dotted and dashed line respectively. Reconstruction of the overlapped first harmonic using CAM results in a poor approximation, since the harmonic envelope of the first harmonic has little similarity with the average of the non-overlapped harmonics. The envelope of the first harmonic is better approximated based on the envelopes from the neighboring non-overlapped notes played on clarinet using HTES.

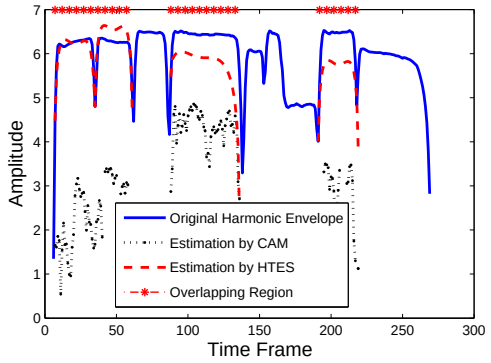

Fig. 5. Original envelopes of the first harmonic from nine notes played by clarinet are plotted in solid blue line. Four notes of 398.4Hz, 397.7Hz, 296.6Hz and 293.3Hz are completely overlapped with four bassoon notes of 132.6Hz, 198.2Hz, 98.2Hz and 146.9Hz.

3. HARMONIC ENVELOPE ESTIMATION

Given a harmonic sound source within an analysis frame (20 ms), the observed frequency-domain signal $Z(m,k)$ at time m and frequency k is the sum of individual signal $X_i(m,k)$ modeled by sinusoidal model in Eq.2.

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

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

where F_i denotes the fundamental frequency and f_s the sampling frequency. $\alpha_i^{h_i}(m)$ is the amplitude parameter and $\phi_i^{h_i}(m)$ is the phase of the $h_i \in \{1, ..., H_i\}$ harmonic of source i. $f_b = f_s/N$ is the frequency resolution and W the analysis window of DFT.

Note Model Construction

Given the non-overlapped harmonics identified by Harmonic Masks, the amplitude from a non-overlapped harmonic h_i of source i is estimated by finding the amplitude $\alpha^{h_i}(m)$ that minimizes Eq.3:

$$\sum_{k \in K_i^{h_i}(m)} \left(|Z(m,k)| - \frac{\alpha^{h_i}(m)}{2} |W(kf_b - h_i F_i(m))|\right)^2 \tag{3}$$

where $|Z(m,k)|$ is the observed spectrogram of the mixture and $k \in K_i^{h_i}(m)$ the set of frequency bins associated with h_i.

The minimization of the above equation is

$$\alpha^{h_i}(m) = \frac{2\sum_{k\in K^{h_i}(m)} |Z(k,m)| \cdot |W(kf_b - h_i F_i(m))|}{\sum_{k\in K^{h_i}(m)} |W(kf_b - h_i F_i(m))|^2} \quad (4)$$

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.4. Assume, h_i is the available non-overlapped harmonic with strongest energy from the note. We re-index the frame indeces by $(x_1, ..., x_N)^T = (1, ..., N)^T$. Fig. 6 shows a plot of a normalized harmonic envelope of a note with length $N = 24$. This envelope is obtained by estimating the harmonic amplitude of the first harmonic of a note played by a clarinet using Eq.4 and normalized by the amplitude of its first frame.

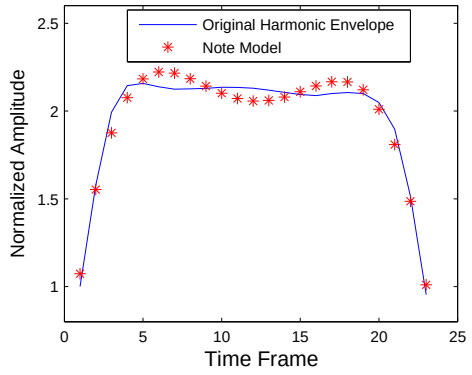

Fig. 6. Original harmonic envelope of a clarinet and its note model: $y = 0.4019 + 0.7804 \times x - 0.1158 \times x^2 + 0.0069 \times x^3 - 0.0001 \times x^4$

Our goal is to exploit the harmonic envelopes from the non-overlapped harmonics to make predictions for the 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:

$$y(x, \mathbf{w}) = \omega_0 + \omega_1 x + \omega_2 x^2 + ... + \omega_M x^M = \sum_{j=0}^{M} \omega_j \theta_j(x) \quad (5)$$

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 $\mathbf{w}$. The values of the coefficients are determined by fitting the polynomial to the harmonic envelope. This can be done by minimizing an *error function* in Eq.6 given by the sum of the squares of the errors between the predictions $y(x_l, w)$ and the corresponding target values r_l.

$$E(w) = \frac{1}{2}\sum_{l=1}^{N} \{y(x_l, w) - r_l\}^2 \quad (6)$$

The solution w^* minimizing Eq. 6 is obtained by:

$$w^* = (\theta^T\theta)^{-1}\theta^T\mathbf{r} \quad (7)$$

where θ is an $N \times (M + 1)$ matrix, whose elements are given by $\theta_{lj} = \theta_j(x_l)$:

$$\theta = \begin{pmatrix} \theta_0(x_1) & \theta_1(x_1) & ... & \theta_M(x_1) \\ \theta_0(x_2) & \theta_1(x_2) & ... & \theta_M(x_2) \\ . & . & . & . \\ . & . & . & . \\ \theta_0(x_N) & \theta_1(x_N) & ... & \theta_M(x_N) \end{pmatrix} \quad (8)$$

For every note that is not completely overlapped, we construct a note model $(\mathbf{w}^*, N)$ for it, where $\mathbf{w}^*$ is the set of polynomial coefficients and N is the length of the note. In Fig. 6, we showed an example of the result of fitting polynomial having order $M = 4$ to a harmonic envelope. Fig.7 showed more note model fitting results to different kind notes played by a clarinet.

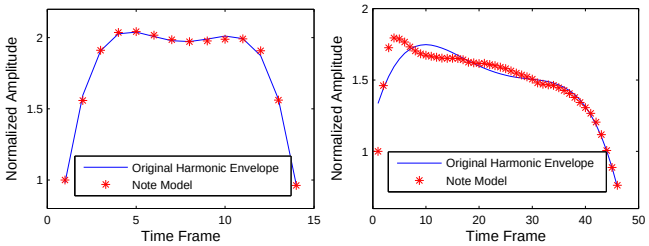

Fig. 7. Note Model fitting to two notes played by clarinet

Envelope Estimation

Given a completely overlapped note of length L, we could approximate its harmonic envelope r^* using an existing note model $(\mathbf{w}^*, N)$ learned from another non-overlapped harmonics of length N by Eq.9

$$r^* = \omega_0^* + \omega_1^* x + \omega_2^* x^2 + ... + \omega_M^* x^M = \sum_{j=0}^{M} \omega_j^* \theta_j(x) \quad (9)$$

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

When a note is not "completely overlapped", we use the note model built from the envelope 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 note model learned from that note to generate a new envelope by Eq.9. These re-estimated envelopes are used in the next stage of our system to estimate the initial amplitude value of the overlapped harmonics.

Fig.5 showed an example of the estimated harmonic envelope by proposed method. The first, second, fourth and eighth notes are completely overlapped by another instrument playing lower pitches. The estimation based on *CAM* (dotted line) is very unstable and different from the original envelope. The dashed line is the envelope estimated by utilizing the note model from the third and last note of the example. Our proposed model produces much better envelope estimates for the completely overlapped notes than the *CAM* does.

4. EXPERIMENT

The proposed system was evaluated on a dataset extracted from 10 Bach chorale recordings performed by local musicians on clarinet, clarinet (trumpet) and violin, totaling about 330 seconds of audio. 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 a Bach chorale. The separation results on violin, clarinet and trumpet are reported because they are extensively "completely overlapped" (two concurrent pitches separated by an octave) with the bass line.

The separation results are measured using source-to-distortion ratio (SDR), source-to-interfering ratio (SIR), and source-to-artifacts ratio (SAR) proposed in [8]. We compare the proposed system to a state-of-art harmonic musical sound separation system [5] (denoted LWW) based on *CAM*. This approach has more difficulty with "completely overlapped" notes. The input to both system is the polyphonic mixture and the fundamental frequency of individual source. We concentrate on the separation of overlapped harmonics itself so the ground truth fundamental frequency of each source is assumed to be given. The estimation of pitch tracks in a polyphonic mixture is another difficult problem. Our previous work on multiple-pitch tracking is described in [9] and has reached promising results.

Fig. 8 showed a real separation example of clarinet from a mixture segment of clarinet and bassoon. There is an improvement (SDR) by 5 dB for 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 irregular envelope. Our proposed system successfully learned the harmonic envelope for clarinet from the non-overlapped harmonics of other notes. Comparing the separated signal by our proposed system to the original signal, we 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.

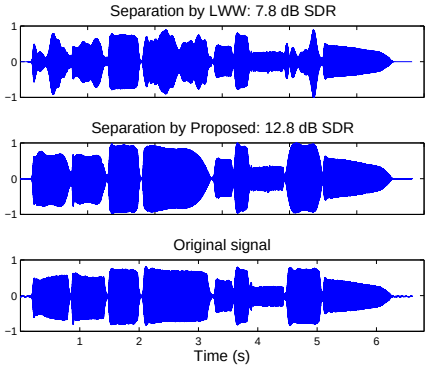

Fig. 8. Separation example of a clarinet from a 6.5 seconds mixture of clarinet and bassoon

Overall separation performance are shown in Table 1. Results are obtained over 50 segments, each 5-to-8 second in length, drawn from the Bach chorales and split at musical phrase boundaries, totalling roughly 330 seconds per instrument. The proposed system improved the separation performance of clarinet and trumpet on SDR and SAR. Specifically, the average improvement compared to LWW on clarinet is about 1.9 dB measured both by SDR and SAR, and 1.1 dB on trumpet. A *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 a polynomial function. More complex models need to be applied to model the harmonic envelope of violin.

5. CONCLUSION

In this paper, we proposed a monaural musical sound separation system that explicitly deals with the "completely overlapped" notes.

Table 1. Performance results of the proposed system and the LWW. Numbers in bold indicate the difference between proposed method and LWW are statistically significant.

Mixtures[a]	SDR		SAR		SIR	
	Proposed	LWW	Proposed	LWW	Proposed	LWW
Clarinet	**12.3**	10.4	**12.3**	10.4	42.3	41.9
Trumpet	**10.7**	9.6	**10.8**	9.6	41.4	39.7
Violin	6.3	6.4	6.4	6.4	44.1	43.7

[a] bassoon as the bass line

Our approach is based on *Harmonic Temporal Envelope Similarity*, a new assumption on instrumental harmonic envelope we observed from the real audio data. 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 authors feel the perceptual quality of the separated signals is improved, as is evidenced by the waveform of the separated signals. This approach works especially well for instruments that have 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.

6. REFERENCES

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

[2] A.P. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proc. ICASSP*, 2001, vol. 5, pp. 3381–3384.

[3] M.R. Every and J.E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Trans. Audio, Speech & Language Processing*, vol. 14, pp. 1845–1856, Sept. 2006.

[4] 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.

[5] Y. Li, J. Woodruff, and D. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Trans. Audio, Speech & Language Processing*, vol. 17, pp. 1361–1371, 2009.

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

[7] J. Han and B. Pardo, "Reconstructing individual monophonic instruments from musical mixtures using scene completion," http://www.cs.northwestern.edu/j̃ha222/paper/han-qual.pdf.

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

[9] Z.Duan, J.Han, and B.Pardo, "Harmonically informed pitch tracking," in *Proc. ISMIR*, 2009.