Semi-supervised Audio Source Separation based on the Iterative Estimation and Extraction of Note Events

Alejandro Delgado Castro[a] and John E. Szymanski[b]
Department of Electronic Engineering, University of York, North Yorkshire, U.K.

Keywords: Audio Source Separation, Note Event Detection, Fundamental Frequency Estimation, Note Event Tracking, Separation of Overlapping Harmonics, Time-domain Subtraction, Semi-supervised Estimation.

Abstract: In this paper, we present an iterative semi-automatic audio source separation process for single-channel polyphonic recordings, where the underlying sources are isolated by clustering a set of note events, which are considered to be single notes or groups of consecutive notes coming from the same source. In every iteration, an automatic process detects the pitch trajectory of the predominant note event in the mixture, and separates its spectral content from the mixed spectrogram. The predominant note event is then transformed back to the time-domain and subtracted from the input mixture. The process repeats using the residual as the new input mixture, until a predefined number of iterations is reached. When the iterative stage is complete, note events are clustered by the end-user to form individual sources. Evaluation is conducted on mixtures of real instruments and compared with a similar approach, revealing an improvement in separation quality.

1 INTRODUCTION

Separating pitched instruments from within polyphonic single-channel mixtures represents a challenging task which has been intensively studied during the last few decades, with direct applications in music information retrieval (MIR), audio coding and compression, content-based analysis, among many others (Zivanovic, 2015). Most of the complexities involved in this process are due to the very rich and non-stationary nature of music, whose evolution over time and frequency creates many regions where the sources overlap (Rafii et al., 2018).

Audio source separation algorithms are based on established signal processing techniques, such as independent subspace analysis (Taghia and Doostari, 2009), non-negative matrix factorization (Bryan and Mysore, 2013), or computational auditory scene analysis (Jang et al., 2003). Estimated sources are extracted using additive synthesis or time-frequency masking, where overlapping content is resolved by sinusoidal modelling (Parsons, 1976), spectral filtering (Every and Szymanski, 2006), common amplitude similarity (Li et al., 2009), amplitude and phase reconstruction (Ponce de León Vázquez and Beltrán Blázquez, 2012), or harmonic bandwidth companding (Zivanovic, 2015). In recent years, deep neural networks have also been explored as a way to introduce machine learning into the separation process (Grais et al., 2017; Chandna et al., 2017).

A common practice followed by various separation approaches is to estimate and extract all underlying sources jointly, relying on a good characterization of their components. One way to characterize audio sources is by tracking their fundamental frequencies across time. However, when pitch trajectories for multiple sources are automatically estimated from the input mixture, their accuracy deteriorates and the complexity of the joint separation approach increases.

On the other hand, an iterative framework in which sources are separated in sections should have several advantages. First, the system only needs to concentrate on separating a small section of audio in every iteration, and second, the number of interacting components should decrease after each section is extracted, reducing the complexity of detecting other sections still present in the mixture.

In this paper, we propose an audio source separation strategy in which the underlying sources are obtained by clustering a set of *note events*, which can be seen as harmonic sounds representing either a single musical note or a group of consecutive notes coming from the same source. These note events are automat-

[a] https://orcid.org/0000-0002-5475-7813
[b] https://orcid.org/0000-0003-2525-654X

Delgado Castro, A. and Szymanski, J.
Semi-supervised Audio Source Separation based on the Iterative Estimation and Extraction of Note Events.
DOI: 10.5220/0007828002730279
In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pages 273-279
ISBN: 978-989-758-378-0

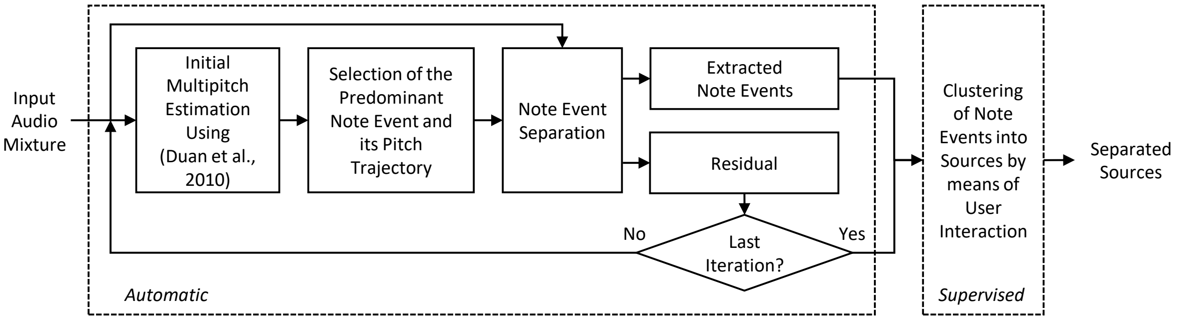
Input Audio Mixture
Initial Multipitch Estimation Using (Duan et al., 2010)
Selection of the Predominant Note Event and its Pitch Trajectory
Note Event Separation
Extracted Note Events
Residual
Last Iteration?
No
Yes
Automatic
Clustering of Note Events into Sources by means of User Interaction
Supervised
Separated Sources

Figure 1: Block diagram of the proposed system showing its two main stages: the automatic detection and separation of note events, and their clustering into individual sources.

ically detected and separated from the input mixture using an iterative approach. Every iteration consists of detecting the pitch trajectory of the predominant note event, separating its spectral content, and extracting its energy from the mixture using subtraction in the time domain. A simplified block diagram of the proposed system is shown in Figure 1.

The rest of the paper is organised as follows. Sections 2 describes the processing stages involved in a single iteration of the system, in which a note event is detected and separated from the input mixture. Section 3 deals with the clustering of note events into sources once the iterative stage is complete. Evaluation is conducted in Section 4 where separation results are compared against the ISSE software package, which is based on a user-informed version of non-negative matrix factorization (NMF) and probabilistic methods. Finally, Section 5 summarizes our conclusions from this work.

2 ITERATIVE STAGE

2.1 Pitch Trajectory of the Predominant Note Event

In every iteration, the input signal is decomposed using the Short-Time Fourier Transform (STFT), without using zero-padding, and the multipitch detector in (Duan et al., 2010) is used to generate the array $\mathbf{P}$ of fundamental frequency estimates, with dimensions $J \times M$, where J is the number of pitch estimates in every frame and M is the number of frames in the decomposition. A salience measure is then assigned to each of these estimates, based on the spectral magnitude summation of their first H partial amplitudes. Considering the m-th frame, the salience of its j-th pitch candidate can be written as:

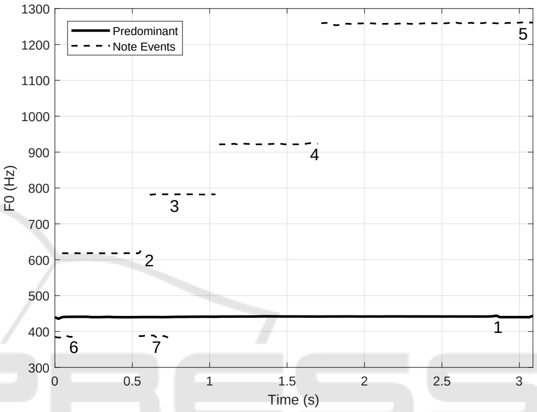

Figure 2: Note events detected in a mixture of viola and clarinet during the first iteration of the system. The viola note has been selected as the predominant event.

$$S_m^j = \sum_{h=1}^{H} \mathbf{X}(m, h f_0^j) \tag{1}$$

where S_m^j is the salience of the j-th pitch candidate in frame m, with fundamental frequency $f_0^j = \mathbf{P}(j,m)$, and $\mathbf{X}(m,f)$ is the magnitude spectrogram of the current input signal. Note events are detected by finding continuous segments of estimates, across all levels of $\mathbf{P}$, for which the change in fundamental frequency between adjacent frames is not higher than one semitone. All detected note events are arranged in a table and their predominances are computed. Considering the k-th note event in the table, existing in level $j = j_k$, starting at frame m_1 and ending at frame m_2, its predominance is defined as follows.

$$S_k^{j_k} = \frac{1}{N_1} \sum_{m=m_1}^{m_2} S_m^{j_k} + \frac{1}{N_2} \left[\frac{m_2 - m_1}{2} \right] \tag{2}$$

where N_1 and N_2 are normalization constants that map the total salience and duration of note events into the range 0 to 1. The note event with the highest predominance is selected as the predominant one. Its pitch

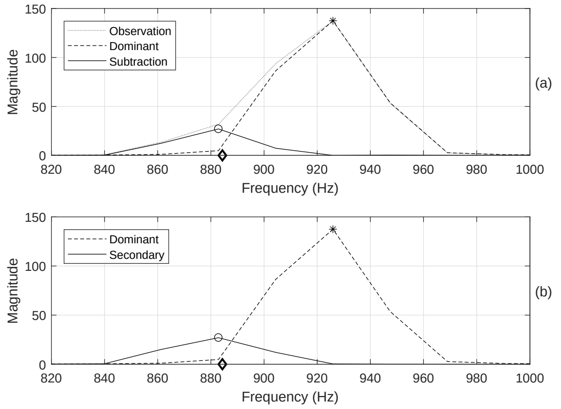

Figure 3: Separation of an overlapping harmonic. (a) Magnitude spectra of the original mixture, dominant component and subtraction. (b) Magnitude spectra of principal and secondary components. A diamond marks the ideal centre frequency of the current harmonic partial.

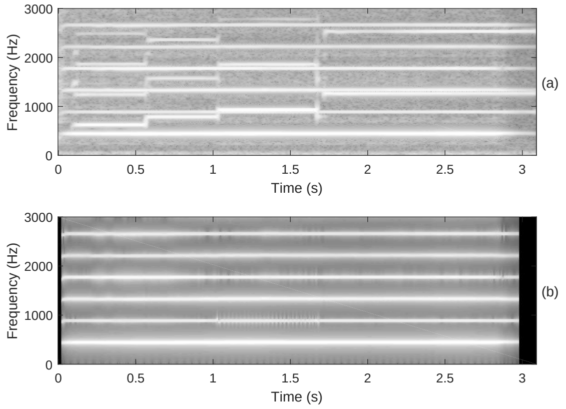

Figure 4: Estimation of the first predominant event in a mixture of viola and clarinet. (a) Spectrogram of the input mixture. (b) Spectrogram of the separated predominant note event (viola A4). In both cases, the frame size is 2048 samples and the hop size is 256 samples.

trajectory, formed by the fundamental frequency estimates assigned to it, is expanded to encompass potential misallocated estimates in adjacent frames. If an adjacent frame has an estimate within a semi-tone of the average pitch of the note event, it is added on to the pitch trajectory of the note event, providing that its salience does not indicate a transition to a different note event. The expanded pitch contour is used to estimate the magnitude spectrogram of the separated predominant note event.

An example is presented in Figure 2 for a mixture of viola and clarinet, where the first plays the note A4, while the second plays the notes D♯5, G5, A♯5, and D♯6. During the first iteration, seven note events are detected in the mixture and the long viola note (event 1) is selected as the predominant one. Notice that note events 6 and 7 do not correspond to real musical notes, they originate from spurious estimates misleadingly generated by the multipitch estimator at this stage, and which are later removed by the system.

2.2 Separated Magnitude Spectrogram of the Predominant Note Event

The pitch trajectory of the predominant note event contains its fundamental frequency estimates and the indexes of the frames in which they are active. It is now possible to find a set of harmonically related partials in every frame, associated with these fundamental frequencies by analysing each magnitude spectrum and finding spectral peaks closest to the ideal harmonic frequencies.

Parameters for each selected spectral peak (centre frequency, absolute magnitude and phase angle) are computed and used to generate a synthetic single-component sinusoidal partial, hereafter referred as the dominant component of the spectral peak. If there is no overlap with other sources, the dominant component can be used to construct the separated magnitude spectrum of the predominant note event in the current frame. However, if the spectral peak also contains contributions from other sources, it is considered as a shared peak, and further processing is required to achieve the separation of its components.

Following the method presented in (Parsons, 1976), the dominant component is subtracted from the shared peak in order to find potential overlapping components. If a significant peak appears in the subtraction, it is treated as energy coming from a different source and its parameters are used to generate a secondary component. Assuming a dual-peak model for the shared peak, in which the observation is a combination of the target harmonic partial plus some other interfering partial, the synthetic component closer to the ideal harmonic frequency, associated with the predominant note event, is selected and used to construct the separated magnitude spectrum.

Figure 3 shows an example of an overlapping partial in the mixture of viola and clarinet previously mentioned, taken from a time frame centred at $t = 1.3$ s. It can be noticed that the dominant component (centred at 925 Hz) is the fundamental harmonic of the clarinet ($f_0 = 922$ Hz), whilst the secondary component (centred at 882 Hz) is the second harmonic of the viola ($f_0 = 442$ Hz). Given that the secondary component is much closer to the ideal position of the second harmonic of the viola, it is selected and used to construct the magnitude spectrum of the separated viola. The magnitude spectrograms of the input mixture

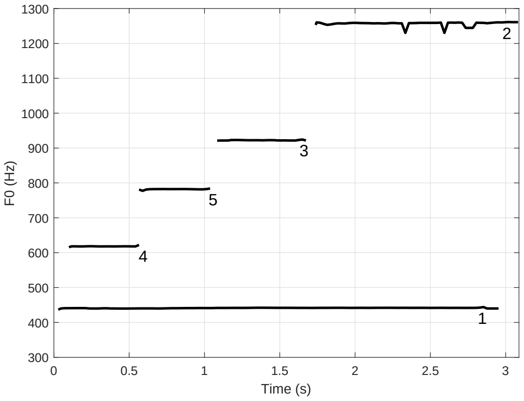

Figure 5: Estimated pitch trajectories of five note events, iteratively extracted from a mixture of viola and clarinet. The numbering of the trajectories follows the extraction order.

and the estimated predominant note event are shown in Figure 4. Notice the separation of the overlapping region between $t = 1.1$ s and $t = 1.7$ s.

2.3 Reconstruction and Subtraction

Time-frequency masking was considered for the extraction of the separated predominant note event from the input mixture, but fitting an appropriate mask proved difficult for note events having low fundamental frequencies. Hence, the extraction of the predominant note event is carried out by reconstructing its separated spectrogram, retaining the original phase information of the mixture, and subtracting it from the input mixture in the time domain. The main advantage of this strategy is that the estimated harmonics of the predominant note event are not significantly distorted by other harmonic partials in the nearby, or by other frequency components associated with other sources.

A residual is also obtain after the subtraction, and it is used as the new input signal for the next iteration. The iterative process continues until a predefined maximum number of iterations is reached.

3 CLUSTERING

At the end of this iterative stage, most of the energy contained in the original mixture should have been allocated within a set of note events, which can be clustered to form individual sources by the end-user, who may use the pitch trajectories of the separated note events as a hint to find an appropriate clustering of the events. The end-user can also listen to each individual note event in order to obtain further guidance. Grouping or instrument identification algorithms could be used at this stage to remove the need for user input, but are not the emphasis of this research. Continuing with the example mixture of viola and clarinet, after five iterations of the system, the final set of estimated pitch trajectories is presented in Figure 5, and their corresponding extracted note events are shown in Figure 6. The end-user is now able to cluster note events 2, 3, 4 and 5 in order to form the separated clarinet, while note event 1 is used to form the separated viola. A comparison between the original and the estimated sources, for the example mixture of viola and clarinet, is presented in Figure 7.

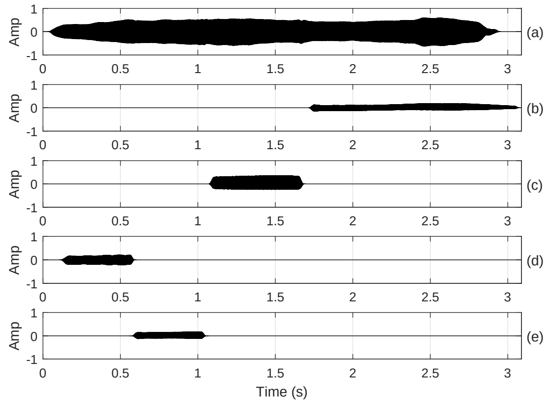

Figure 6: Extracted note events from a mixture of viola and clarinet. (a) Viola A4, (b) Clarinet D♯6, (c) Clarinet A♯5, (d) Clarinet D♯5, and (e) Clarinet G5.

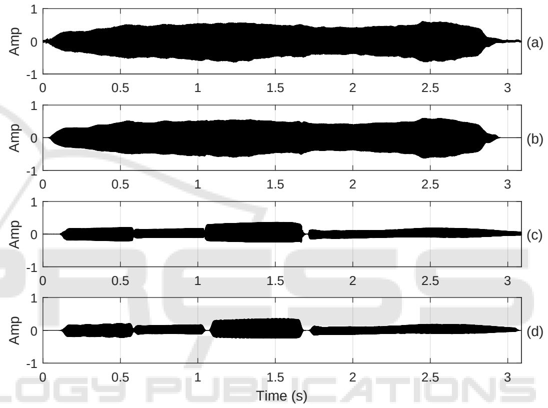

Figure 7: Original and estimated sources from a mixture of viola and clarinet. (a) Original viola, (b) Estimated viola, (c) Original clarinet, and (d) Estimated clarinet.

4 EVALUATION

Separation performance is evaluated in three different experiments, where the proposed algorithm is applied to a number of audio mixtures. The first two experiments consider audio mixtures consisting solely of pitched sources, while the third one introduces one percussive source. Four pitched instruments are studied (violin, clarinet, tenor saxophone and bassoon), taken from excerpts of the Bach10 database (Duan et al., 2010). The percussive source consists of a synthesized sequence of snare drums and cymbals. These test recordings are available online[1].

The quality of the separation is assessed by measuring the source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifacts ratio (SAR), as defined in (Vincent et al., 2006), where each estimated source $\hat{x}_i$ is decomposed as follows.

$$\hat{x}_i = x_{\text{target}} + e_{\text{interference}} + e_{\text{noise}} + e_{\text{artifacts}} \tag{3}$$

where $x_{\text{target}} = f(x_i)$ is a version of the true source x_i modified by some allowed distortion $f(\cdot)$, and where $e_{\text{interference}}$, e_{noise}, and $e_{\text{artifacts}}$ are the interferences, noise and artifacts error terms, respectively. These terms should represent the part of $\hat{x}_i$ perceived as coming from x_i, from other unwanted sources, from sensor noise, and from other causes.

The aforementioned objective measures assign equal weights to all error terms, which means that all types of distortions contribute equally to the overall quality of the extracted source (Cano et al., 2016). A set of MATLAB® functions, created by Févotte et. al. and referred as *BSS_Eval Toolbox*[2], is available online and can be used to calculate these objective measures (Févotte et al., 2005).

Separation results are averaged out in every experiment and compared with another semi-supervised approach, known as the Interactive Source Separation Editor (ISSE) (Bryan and Mysore, 2013), where the end-user provides annotations in order to constrain, regularize, or otherwise inform the algorithm. These annotations are introduced at the beginning of the process by highlighting relevant sections on the input spectrogram, while the separation of the sources is obtained by an implementation of the NMF approach. Within the existing user-assisted audio source separation methods, ISSE constitutes a representative example that has the additional advantage of being open-source and freely available[3].

Oracle estimates are also calculated in every mixture, according to Vincent et. al. (Vincent et al., 2007), and their averages are presented in every experiment as a reference. In theory, they represent the highest achievable results that a time-frequency masking-based separation method can obtain. A set of MATLAB® functions is also available online[4] and can be used to calculate these estimates, in particular, the function *bss_nearopt_monomask*, which generates near-optimal time-frequency masks using the STFT with a sine window (Vincent and Plumbley, 2007).

The proposed iterative estimation/separation system (IES) is applied with a frame size of 2048 samples, 87.5% overlap, a Hanning window, and $H = 5$ partials for the salience measurement. In every frame, the maximum number of extracted harmonic partials is set to 30 in order to capture most of the energy associated with the selected note event. The maximum number of note events to be extracted from within every mixture is set to 45. The ISSE, on the other hand, is applied to every mixture using the recommended settings (frame size of 4096 samples and 50 basis vectors per source), while the annotations are introduced to extract one source at a time.

4.1 Two Harmonic Sources

In this experiment, a set of 18 audio mixtures with polyphony 2 are considered. Overall, the number of notes being played is 279, with fundamental frequencies spanning from F2 (86 Hz) to F♯5 (750 Hz). Separation results are presented in Table 1, where the IES system shows an average improvement of 25% in SDR over the ISSE algorithm.

Table 1: Separation Performance in Audio Mixtures with Polyphony 2 (Harmonic Instruments).

Method	Separation Performance (dB)		
	SDR	SIR	SAR
IES	12.87	19.93	14.66
ISSE	9.35	12.76	15.05
Oracle	18.16	26.96	19.01

Although the ISSE seems to generate slightly less artifacts, the separated sources also exhibit higher levels of interference, suggesting that the annotations are not providing enough information to completely characterize each individual source. This problem is partially solved in IES by assuming that the underlying sources are harmonic, which provides a simple but effective way to identify their frequency components based on the knowledge of their fundamental frequencies. The proposed dual-peak model provides a sharper separation of shared harmonics, which also reduces interference among the separated sources.

4.2 Three Harmonic Sources

A different set of 12 audio mixtures with polyphony 3 are now considered, where the number of notes being

[1] http://www-users.york.ac.uk/ adc533/download

[2] http://bass-db.gforge.inria.fr/bss_eval/

[3] http://isse.sourceforge.net/

[4] http://bass-db.gforge.inria.fr/bss_oracle/

Table 2: Separation Performance in Audio Mixtures with Polyphony 3 (Harmonic Instruments).

Method	Separation Performance (dB)		
	SDR	SIR	SAR
IES	8.81	15.69	10.63
ISSE	7.31	9.34	13.17
Oracle	14.04	23.20	14.79

played is 386, and their fundamental frequencies are in the range F3 (175 Hz) to F♯5 (750 Hz). Results for this experiment are presented in Table 2.

The incorporation of a third source represents a reduction of the separation quality, as can be observed for both algorithms. A higher number of simultaneous sources means additional difficulties in providing good annotations for the sources, reducing the overall performance of ISSE, but it also means additional problems during the separation of overlapping harmonics, which affects IES quality. However, the higher number of note events in the mixture and the proximity of their frequency components are causing a higher reduction in the separation performance of IES, in comparison with the previous experiment.

Octave-related notes, which are present in some of the mixtures analysed in this experiment, introduce an additional challenge for both algorithms and affect the separation performance. The IES system is able to detect the pitch trajectories of many octave-related notes, however, an accurate separation of the original note events is not possible, since the amplitudes of their harmonic partials cannot be correctly estimated from the mixed spectrogram. Similarly, the ISSE system also has problems interpreting overlaps between annotations of different sources and tends to allocate most of the shared energy into only one of the sources.

4.3 Two Harmonic and One Percussive Sources

The third experiment considers the same set of mixtures used in Section 4.2, but the third pitched instrument (tenor saxophone) is replaced with a percussive source. A total of 239 harmonic notes are still present, with fundamental frequencies in the range A3 (220 Hz) to F♯5 (750 Hz), and several hundred new percussive events are introduced. Results for this experiment are presented in Table 3.

The IES method presented here is designed to detect harmonic content, consequently the percussive output is contained in a residual signal together with other non-harmonic content. In the case of ISSE, the percussion is instead extracted first by exploiting additional user-provided annotations of solo percussive regions of the spectrogram.

Table 3: Separation Performance in Audio Mixtures with Polyphony 3 (Harmonic and Percussive Instruments).

Method	Separation Performance (dB)		
	SDR	SIR	SAR
IES	11.98	18.36	13.49
ISSE	11.32	16.14	14.26
Oracle	14.86	24.60	15.65

In this experiment, both algorithms show similar separation quality, with the IES approach still showing slightly less interference in the separated sources, while the ISSE approach introduces slightly less artifacts. In this specific example, the percussive source does not affect the detection of note events within the IES system but, more generally, low energy percussive effects might impact on the detection of musical notes with a fundamental frequency below 200 Hz.

An important advantage of IES over ISSE is that it allows end-user interaction during the final stage of the process (clustering of note events), which seems to be more effective than using it at the beginning of the separation, as in the case of the ISSE process. From the user perspective, listening to separated events and grouping them into individual sources is far easier than recognising harmonic structures and estimating frequencies from within the spectrogram of a complex audio mixture.

5 CONCLUSIONS

In this paper, a novel semi-supervised approach for single-channel audio source separation was introduced, based on the iterative estimation and extraction of note events, and their subsequent clustering into separated sources by end-user interaction during the final stage of the process. Direct subtraction in the time domain is used here during the separation of each note event, which provides a softer way of extracting its estimated spectral energy from within the mixture and reduces the levels of interference between the separated sources.

After evaluation on a set of test mixtures with polyphonies 2 and 3, the proposed system outperformed the ISSE NMF-based approach, in which end-user interaction is used at the beginning of the separation process. Positive separation results were also obtained by the IES system for audio mixtures with polyphony three including percussive effects, despite the complexities of performing pitch tracking in the presence of percussive sounds. Finally, grouping separated note events into sources was found to be more effective than recognising structures from within the spectrogram of a complex audio mixture.

Further work will be conducted with the aim of allowing the separation of notes in octave relation, and improving the separation of low-pitched notes. Different approaches to automate the clustering of note events into sources will also be explored, as a way to deliver a fully-automated source separation system that could be compared with other unsupervised algorithms based on machine learning.

ACKNOWLEDGEMENTS

The authors would like to thank the University of Costa Rica and the Costa Rican Ministry of Science, Technology and Telecommunications for their support in founding this research.

REFERENCES

Bryan, N. J. and Mysore, G. J. (2013). Interactive refinement of supervised and semi-supervised sound source separation estimates. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 883–887.

Cano, E., Fitzgerald, D., and Brandenburg, K. (2016). Evaluation of quality of sound source separation algorithms: human perception vs quantitative metrics. In *Proceedings of the 24th IEEE European Signal Processing Conference*, number 1, pages 1758–1762.

Chandna, P., Miron, M., Janer, J., and Gómez, E. (2017). Monoaural audio source separation using deep convolutional neural networks. In *Proceedings of the 13th International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266.

Duan, Z., Pardo, B., and Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech and Language Processing*, 18(8):2121–2133.

Every, M. R. and Szymanski, J. E. (2006). Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1845–1856.

Févotte, C., Gribonval, R., and Vincent, E. (2005). BSS EVAL toolbox user guide. Technical Report 1706, Institut de Recherche en Informatique et Systèmes Aléatoires.

Grais, E. M., Roma, G., Simpson, A., and Plumbley, M. D. (2017). Two-stage single-channel audio source separation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(9):1469–1479.

Jang, G. J., Lee, T. W., and Oh, Y. H. (2003). Single-channel signal separation using time-domain basis functions. *IEEE Signal Processing Letters*, 10(6):168–171.

Li, Y., Woodruff, J., and Wang, D. (2009). Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Transactions on Audio, Speech and Language Processing*, 17(7):1361–1371.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60(1976):911.

Ponce de León Vázquez, J. and Beltrán Blázquez, J. R. (2012). Blind separation of overlapping partials in harmonic musical notes using amplitude and phase reconstruction. *EURASIP Journal on Advances in Signal Processing*, (223):1–16.

Rafii, Z., Liutkus, A., Stoter, F. R., Mimilakis, S. I., Fitzgerald, D., and Pardo, B. (2018). An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(8):1307–1335.

Taghia, J. and Doostari, M. A. (2009). Subband-based single-channel source separation of instantaneous audio mixtures. *World Applied Sciences Journal*, 6(6):784–792.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469.

Vincent, E., Gribonval, R., and Plumbley, M. D. (2007). Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8):1933–1950.

Vincent, E. and Plumbley, M. D. (2007). BSS ORACLE toolbox version 2.1 user guide. Technical report.

Zivanovic, M. (2015). Harmonic bandwidth companding for separation of overlapping harmonics in pitched signals. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(5):898–908.