Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

On the Importance of Temporal Context in Proximity Kernels: A Vocal Separation Case Study

Delia Fano Yela[1], Sebastian Ewert[1], Derry FitzGerald[2], and Mark Sandler[1]

[1] *Queen Mary University of London*
[2] *Cork School of Music, Cork Institute of Technology*

Correspondence should be addressed to Delia Fano Yela (d.fanoyela@qmul.ac.uk)

ABSTRACT

Musical source separation methods exploit source-specific spectral characteristics to facilitate the decomposition process. Kernel Additive Modelling (KAM) models a source applying robust statistics to time-frequency bins as specified by a source-specific kernel, a function defining similarity between bins. Kernels in existing approaches are typically defined using metrics between single time frames. In the presence of noise and other sound sources information from a single-frame, however, turns out to be unreliable and often incorrect frames are selected as similar. In this paper, we incorporate a temporal context into the kernel to provide additional information stabilizing the similarity search. Evaluated in the context of vocal separation, our simple extension led to a considerable improvement in separation quality compared to previous kernels.

Introduction

Music recordings typically comprise a mixture of different sound sources, corresponding to musical instruments such as a guitar, drums or vocals. Many applications including up-mixing, automatic transcription or musical feature extraction require or benefit from sources being isolated or enhanced from the rest of the mixture. Even if the signal only contains a single sound source, one may be interested in separating different aspects of the source for further analysis, for example, to differentiate transients related to onsets from the pitched signal components. In this context, various models and techniques have been proposed. Depending on the use case, a major goal is to find characteristics helping with the definition, identification and separation of individual sources. Such characteristics can include various acoustical or perceptual aspects, including the typical behaviour of a source in time (e.g.: vibrato [1, 2], continuity in activity [3, 4], repetitiveness of patterns [5, 6]) or spectral characteristics (e.g.: broadband vs harmonic energy distribution [7]). These properties are often either modelled explicitly [8] or are learned from data [9].

A particularly successful family of techniques is based on Non-Negative Matrix Factorization (NMF) [10], where a time-frequency representation of a signal is modelled as a product of two matrices. The first matrix captures the spectral properties of the signal in its columns, each often referred to as a spectral *template*. The corresponding rows (or *activations*) in the second matrix determine when and how strong each template is present in the signal. As applying the original NMF ap-

proach [10] to musical data often does not yield useful results [11], most of the state-of-the-art source separation methods are based on NMF variants [12] incorporating additional information about the source in the form of spectral constraints [3, 4], user-assisted annotations [13] or score information [14]. Compared to previous approaches such as Independent Subspace Analysis, NMF relies slightly less on the assumption of statistical independence among sources [15]. However, as a severe limitation in NMF, the results obtained typically strongly depend on how well the spectral templates reflect the actual properties in a given recording. In particular, changes with respect to the instrument or recording conditions can lead to a drastic decrease in separation performance.

The recently proposed Kernel Additive Modelling (KAM) [16] takes a different approach. Assuming that several sources overlap in a specific bin in a time-frequency representation, the idea is to reconstruct the magnitude for a given source in that bin by analysing the values in other bins, in which the source is likely to assume similar values. The similarity relation between bins is specified by a source-specific kernel, which defines for each pair of bins whether they are to be called similar or not – typically using some sort of underlying metric in the background. The goal in KAM is to design a source-specific proximity kernel that indicates, given a specific bin, where to look for similar bins in a time-frequency representation of the signal. Once a set of similar bins has been identified for a source, the contribution of the remaining sources can be regarded as outliers – assuming that not all sources have the exact same kernel function. This way, KAM produces an estimate of the magnitude in a bin for a specific source by applying robust statistics (typically the median) across the similar bins.

For vocal separation, existing KAM instances [5, 6] considered the accompaniment to be far more stationary and repetitive than the vocal source. This means that there are many time frames containing the same (or similar) accompaniment but not many with the same voice content. Further, it is implicitly assumed that the energy contribution of the accompaniment is higher than that of the vocal source. In line with this reasoning, the kernel proposed in [5] is a function that, based on the Euclidean distance, returns the K frames most similar to a given frame. More precisely, in this case, a bin is considered similar to a second bin if both have the same centre frequency and the frame number for the second is among the K most similar frames. Based on the above assumptions, the voice thus can be regarded as an outlier and can be eliminated using median filtering across the similar bins.

Even though this kernel can exploit some of the source's regularities, its simplicity leads to some drawbacks. To illustrate this, let's consider a recording containing two instrumental solo sections for a guitar and a piano. Depending on the recording conditions, the sustain part for both instruments can have a similar energy distribution in frequency direction (playing the same musical pitch). As a consequence, a frame-wise kernel based on the Euclidean distance sometimes fails to identify the actual dissimilarity between frames and can confuse a guitar frame with a piano frame. Such issues are even more pronounced if an instrument has variable timbre, for example due to the use of effects. This mix-up can lead to an unexpected energy distribution for an instrument in the separation result.

Using only a single frame, such issues are difficult to resolve. However, by taking the temporal context of a frame into account, we obtain more information about which frames are actually similar to each other. For example, using a larger temporal context, the similarity measure might take a frame containing the onset into account, which can be very discriminative for an instrument. Also, the temporal context might even be large enough to pick up some basic information about the musical context and, assuming the different instruments play different note patterns, we can use this low-level musical context as additional guidance to find similar frames for a given instrument.

Based on this simple idea, we propose in this paper to modify existing kernels by introducing a temporal context. Basically, given a frame we aim to find similar frames for, we include the preceding and succeeding frames in the similarity function underlying our kernel. Effectively, that means we measure similarity based on entire groups of frames instead of single frames. The size of the temporal context is chosen large enough to give some rough indication of local musical patterns. Re-using the previous guitar-piano example and assuming that the current frame is in the guitar solo, the group of frames centered around this frame might span a few notes. Now, when looking for similar segments, we can take this local note constellation to some degree into account, which potentially aids in differentiating between similar timbres. In particular, unless the solo piano

section contains the same sequence of notes played in the same fashion, the guitar will not be mistaken for the piano.

The remainder of the paper is organized as follows. In Section 2 we give a brief overview of related work, followed by Section 3, where we describe the details of our proposed extension. Next, in Section 4 we report on experiments indicating the level of improvement resulting from our extension. Finally, we conclude in Section 5 with an outlook to possible research directions.

Related Work

Methods for musical source separation typically incorporate various types of prior knowledge about the individual sound sources or the mixing process. In a vocal separation scenario, one can exploit various properties of the singing voice and of the background or accompaniment. Some methods start by differentiating between vocal and non-vocal regions. For example, the method presented in [17] uses a priori knowledge about non-vocal segments to learn an accompaniment model based on Probabilistic Latent Component Analysis (PLCA) and then fixes the accompaniment to learn the vocal source. The methods presented in [18, 19] employ Mel Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Models (GMM) to first differentiate between vocal and non-vocal regions and then use the resulting information to train a Bayesian model for the accompaniment [18]. Using similar pre-processing steps, the approach introduced in [19] extracts the vocal pitch contour using a predominant pitch estimator on the vocal segments and performs separation through binary masking. Similarly, [20] uses a predominant pitch estimator based on a Hidden Markov Model (HMM) to extract the vocal pitch contour through spectral subtraction in voiced segments and GMMs to identify and separate the unvoiced consonants.

Another large body of work is based on Deep Neural Networks (DNNs) that are typically trained on the magnitude spectrogram of the mixture to predict either a time-frequency mask describing the energy distribution of a source relative to the other sources [21] or the source spectrogram directly [22, 9]. In [9] the authors employ a DNN to extract a spectrogram for each source using multi-channel recordings as input; the parameters of a multi-channel Wiener filter are estimated using an iterative Expectation-Maximization (EM) algorithm. While most state-of-the-art methods for vocal separation employ some variant of DNNs, these methods are typically trained for specific combinations of instruments or instrument groups, which limits their flexibility and adaptivity in practice. Further, the performance of these techniques depends strongly on the quality of the training data.

Instead of training a model, various methods target the inherent properties of the sources directly and use their differences to distinguish them in the separation process. For instance, the repetition of patterns is a core characteristic of popular music. In this context, some methods [23, 24] employ Robust Principal Component Analysis (RPCA) to decompose the spectrogram into a low-rank and a sparse matrix, and argue that these can be associated with the accompaniment and vocals, respectively. The method presented in [24] takes this idea a step further by introducing vocal activity information into the standard RPCA algorithm.

For cases in which the background music can be considered to be repetitive, the method REPET [6] identifies the repetition period of the musical pattern, models the repeating segment and, through the use of robust statistics and soft-masking, extracts the repeating pattern associated with the background music. Even though it has proven successful in a variety of contexts, it is limited to only one repeating pattern and is unable to further differentiate individual sound sources. The method [6], as well as others relying on a similarity measure such as [5, 25], can be considered as instances of the more general KAM framework [16] described in the introduction. The approach presented in this paper is also a member of the KAM family of methods as it represents an extension to the method described in [5].

Proposed Method

To describe our approach, we follow the notation used in [5]. The method can be regarded as an instance of KAM using only one iteration of the Kernel Backfitting procedure described in [16].

More precisely, let $C \in \mathbb{C}^{M \times N}$ be the spectrogram of a music recording containing a mixture of a vocal and an accompaniment track, where M is the number of frequency bins and N the number of time frames, see Fig. 1. Further, let $X \in \mathbb{R}^{M \times N}$ be the corresponding magnitude spectrogram. As a first step, the vocal separation algorithm proposed in [5] computes for

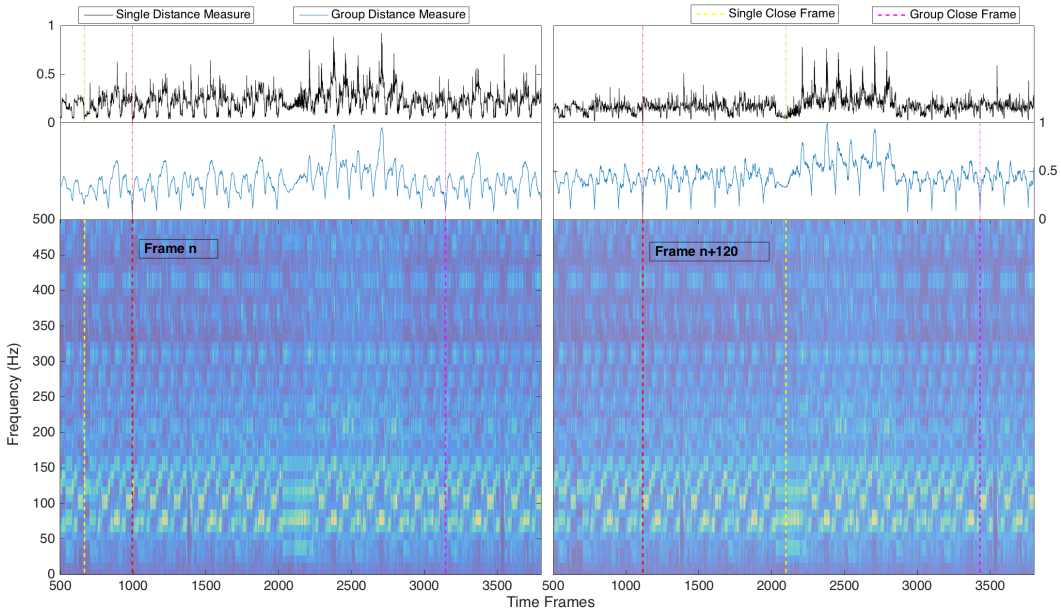

Fig. 1: Comparison between single and group distance measure: given a spectrogram of the mixture (shown in each half of the figure), the plots show the distance between a specific frame ($k = 1000$ on the left and $k = 1120$ on the right, see red vertical lines) and the remaining frames. The single-frame and group distance measure are plotted in black and blue, respectively. The frames closest with respect to these distances are indicated using yellow (single-frame) and magenta (group) vertical lines.

each pair of frames $(k,\ell) \in \{1,\ldots,N\} \times \{1,\ldots,N\}$ the Euclidean distance between the two corresponding columns in X:

$$D_{k,\ell} = \sum_{m=1}^{M} (X_{m,k} - X_{m,\ell})^2.$$

In Fig. 1 the k-th row of D is plotted in black, with $k = 1000$ in the left and $k = 1120$ in the right half of the figure. The spectrogram of the mixture to be processed is shown in each half of the figure and the two frames are indicated by a vertical red line.

To obtain a list of P frames being closest to a given frame k, the symmetric matrix D is first sorted individually in each row. By keeping track of which frame index belongs to which entry in the sorted matrix D, we can create for each k a matrix $A^k \in \mathbb{R}^{M \times P}$ in which the P columns contain the P closest frames, i.e. a specific subset of frames taken from X. This process is illustrated in the left part of Fig. 1, where the most similar frame (indicated by a yellow vertical line) is found using the similarity values in D. For the upper part of Fig. 2, this process was repeated until for a given frame (shown in Fig. 2a) the $P = 10$ closest frames were found (shown in Fig. 2b-k), i.e. these 10 frames represent the first ten columns of A^k.

Assuming that the energy in each frame and hence the Euclidean similarity measure are dominated by the accompaniment, we now have P frames similar to frame k that only differ in terms of the vocal part (following above assumptions). In particular, the vocal part leads to outliers and we want to extract the commonalities between the frames in A^k. To this end, the method in [5] employs the median filter, which is invariant against outliers (up to 50 percent) and belongs to the class of operator used in robust statistics. More precisely, we define the estimated magnitude spectrogram $Y \in \mathbb{R}^{M \times N}$ as follows:

$$Y_{m,k} := \text{median}(A^k_{m,1},\ldots,A^k_{m,P})$$

To extract both magnitude and vocals from the mixture, we create a mask (similar to a Wiener filter) as in [5]. More precisely, we measure the distance between the mixture X and the accompaniment estimate Y after a logarithmic compression (with the logarithm leading

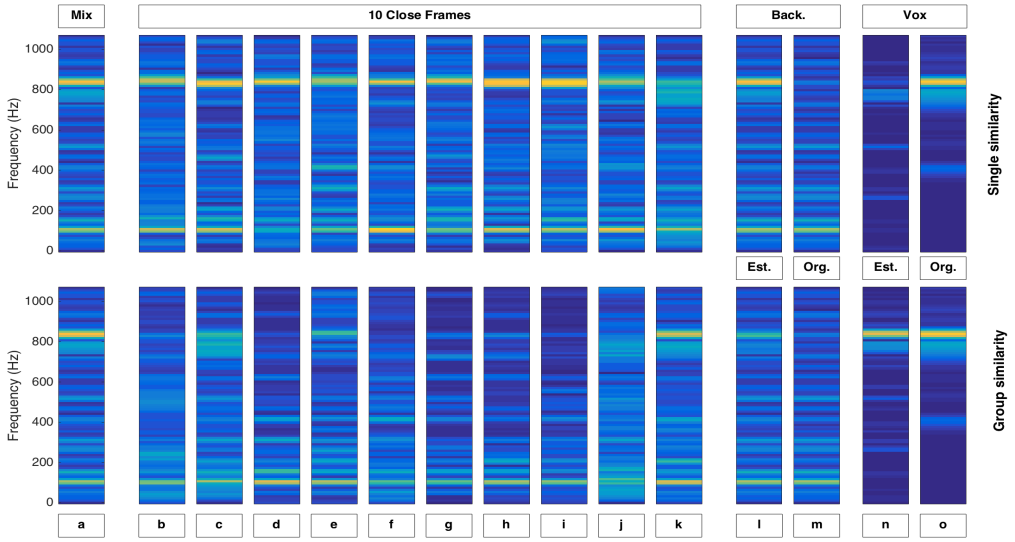

Fig. 2: Given a mixture frame (a), the first ten close frames (b)-(k) identified using single and group similarity yield a background music (l) and vocal (n) approximations of the clean sources, (m) and (o) respectively.

to a perceptually more meaningful distance [5]) and employ this distance in a Gaussian radial basis function to obtain a mask $W \in [0,1]^{M\times N}$:

$$W_{m,n} = \exp\left(-\frac{(\log(X_{m,n}) - \log(Y_{m,n}))^2}{2\lambda^2}\right),$$

where λ is a parameter to additionally compress the log-distances non-linearly. In the following, we set λ=1. The complex spectrograms for the accompaniment $B \in \mathbb{C}^{M\times N}$ and vocals $V \in \mathbb{C}^{M\times N}$ can then be estimated by applying the soft masks W and $(1-W)$ to the original mixture spectrogram C using an element-wise multiplication $\odot$, respectively:

$$B = W \odot C$$

$$V = (1-W) \odot C$$

With this framework in place, our extension can be explained in a simple, straightforward way. In particular, we keep the same basic procedure but replace D with a new pairwise distance matrix $\widetilde{D}$, taking a temporal context for each pair of frames additionally into account. More precisely, we define:

$$\widetilde{D}_{k,\ell} = \sum_{m=1}^{M} \sum_{r=-R}^{R} (X_{m,k+r} - X_{m,\ell+r})^2,$$

where $R \in \mathbb{N}$ is a radius in number of frames defining the extend of our temporal context. This way, we do not compare single frames anymore but whole groups of frames (resembling concepts used in the context of musical structure analysis [26]).

Overall, this change in the distance matrix is a rather small extension – however, we found this small change to have a reasonably strong impact on the separation result. To illustrate our findings, we discuss Figures 1 and 2 now in more detail. We start with the first row in Fig. 2. Fig. 2a shows a frame taken from the input X that we wish to process. As we can see, the frame contains two strong partials, one corresponding to the background music (100Hz, compare also Fig. 2m for the ground truth) and another one related to the vocal source (830Hz, compare Fig. 2o). Using the single-frame distance D, we obtain results equivalent to the KAM-based baseline [5]: the 10 closest frames, taken from A^k, are shown in Figs. 2b-k. As we can see, due to the strong vocal activity in the input, the single frame distance used in D is heavily influenced by the vocal partials and frames are selected in Figs. 2b-k that are also dominated by similar vocal activity. As a consequence, even applying the median filter to these frames in A^k does not help with the identification of the vocal partial as an outlier – simply because the vocal partial occurs in every frame. Comparing the median filtered result for the accompaniment (Fig. 2ℓ, computed using $P = 100$) with the ground truth (Fig. 2m), we observe that the vocal partial remains intact and the separation

was ineffective – compare also Fig. 2n, which contains the vocal estimate which hardly contains energy. We have identified this problem to appear consistently in scenarios with a low Signal-to-Noise Ratio (SNR), taking the signal as the source we wish to isolate and the remaining sound sources as noise. This behavior represents a considerable drawback in the current kernel design proposed in KAM.

Once we introduce a temporal context in the distance function, the importance of the frame to be filtered is lowered while the importance of having a similar neighbourhood is increased. To see this, we now look at the second row in Fig. 2, where the results are shown using our extension. Taking a look at the 10 closest frames (Figs. 2b-k), we notice a higher diversity among them compared to the previous scenario – however, the vocals are a lot less dominant while the accompaniment is more prominent. Looking at the result after the median filter (Fig. 2ℓ, computed using $P = 100$), we see that there is still some vocal energy left but, compared to the single-frame distance, the vocals are much more suppressed and the result is considerably closer to the ground truth (Fig. 2m). The improvement is also clearly visible in the separation result for the vocals (compare Fig. 2n and o).

It is also interesting to compare the two distances directly in the form of the matrices D and $\widetilde{D}$. Fig. 1 shows the row for frame k in D (plotted in black) and in $\widetilde{D}$ (plotted in blue), with $k = 1000$ in the left part of the figure and $k = 1120$ for the right part. As we can see, the single frame distance is much more noisy compared to the one with temporal context. Also, peaks indicating a low distance (i.e. high similarity) are much clearer for the curve using a temporal context – this is particularly visible in the right half of the figure where many spurious peaks can be found in the single frame distance. This overall change in qualitative behaviour also influences which frames are selected as the most similar frames. In particular, the yellow and magenta lines in Fig. 1 indicate the most similar frame found using the single frame and group of frames distance, respectively.

Comparing the yellow and magenta position in the spectrogram on the left, we see that both indicate a frame with a low distance that additionally makes sense musically as both happen at the end of a similar note constellation. In the right half of the figure ($k = 1120$), however, we can notice that the frame selected via the single frame distance is in a completely different section of the song. Zooming in, we find that the magnitude values are indeed similar which explains this selection but, being in a different section of the song, there are various subtle differences in that frame leading to additional difficulties for the median filter. Looking at the distance values using the temporal context, we see that around the hit for the single-frame distance, the distance values are here quite high, which indicates a musical dissimilarity to the pattern around location $k = 1120$.

Experiments

To quantitatively compare our proposed extension with the baseline [5], we employed the Demixing Secrets Dataset 100 (DSD100) as also used in the 2016 Signal Separation Evaluation Campaign (SiSEC) [27]. The dataset contains 100 different songs of various genres, all of them being polyphonic, mixed in stereo, with a sampling frequency of 44.1 kHz and 30s long. In order to asses the separation quality, we used the toolkit available in the SiSEC website, which employs the BSS Eval toolbox 3.0 [28] to calculate the Signal to Distortion Ratio (SDR), Source Image to Spatial Distortion Ratio (ISR), Source to Interference Ratio (SIR) and the Source to Artifacts Ratio (SAR).

For this evaluation, we have implemented (using an FFT size of 4096 and a hopsize of 2048 samples) an instance of KAM for vocal separation following [5] and introduced a temporal context in the proximity kernel as described in Section 3. The number of frames R specifying the temporal context is a parameter of our approach. In principle, this parameter could be adapted for every frame based on musical knowledge, for example, based on segmentation information or pitch tracking data, which would render the method more flexible and adjustable to musical changes in the signal. However, for this paper, we chose to use a fixed setting for the radius R. To find a suitable value, we conducted a simple parameter sweep, whose results are shown in Fig. 3. The figure shows the averaged SDR values for both vocal and accompaniment separation using the proposed extension for different radius values. We can observe an overall trend in Fig. 3 shared by both vocal and accompaniment separation, where the biggest difference in SDR value is between a zero radius (the baseline method) and the other values taking a temporal context into account. In addition, we see

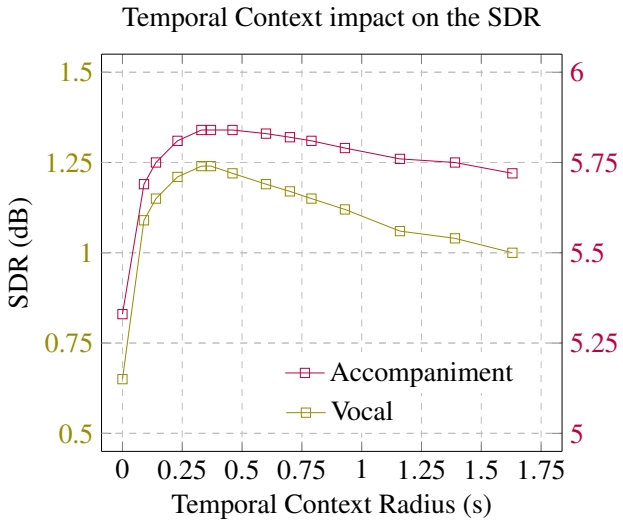

Fig. 3: SDR results for the proposed extension with different temporal contexts for the DSD100 dataset.

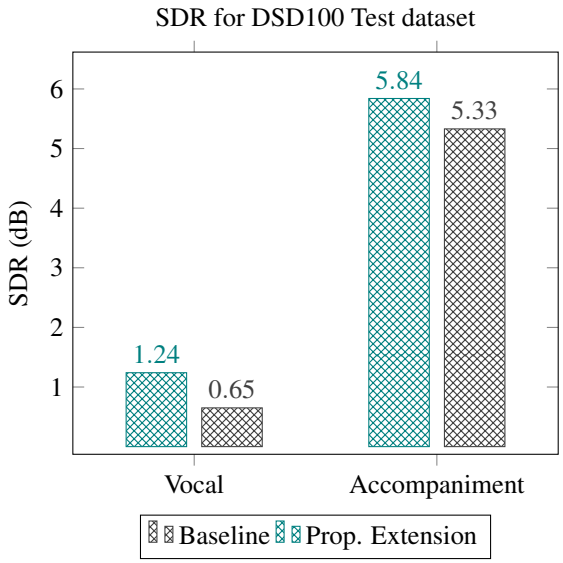

Fig. 4: SDR results for the baseline method and the proposed extension tested on the DSD100 dataset.

that the highest SDR values are achieved for a temporal context of around 1 second (radius values between 0.25 and 0.6 seconds), which can be considered wide enough to capture some simple musical patterns. If the radius is increased, the musical information within the temporal context grows and we observe a slight decrease of the SDR. For this reason, we chose to fix the radius to 372ms.

Using this fixed value for R, Figure 4 shows the SDR values comparing our proposed method to our baseline [5] (i.e. using single frame distances) in more detail. On the SiSEC dataset, our proposed method consistently outperforms the baseline and improves the results by about 0.5dB SDR on average for both vocal and accompaniment separation. Given the simplicity of our extension, this is quite considerable.

Overall, the results are encouraging for such a simple unsupervised method that requires no prior training. Even though there is still room for improvement, introducing temporal context in the similarity search has shown clear advantages.

Conclusion

We presented a simple approach to improve the similarity search in proximity kernels as used in the KAM framework for source separation. We motivate the need for introducing a temporal context in kernels by analyzing different scenarios where the similarity search would fail otherwise. The results obtained show an improvement in separation performance compared to the baseline on the DSD100 dataset used in SiSEC 2016. Our results indicate that our extension to the similarity measure temporally stabilises the source estimates and improves the separation performance over the baseline algorithm. Possible future directions include a more extensive study of different approaches to adaptively set the length of the temporal context, taking source-specific characteristics into account on a frame-by-frame level.

Acknowledgement: This work was funded by EPSRC grant EP/L019981/1.

References

[1] J. Driedger, S. Balke, S. Ewert, and M. Müller, "Template-based vibrato analysis in music signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2016, pp. 239–245.

[2] J. Driedger and M. Müller, "Extracting singing voice from music recordings by cascading audio decomposition techniques," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 126–130.

[3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity

and sparseness criteria," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[4] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, 2010.

[5] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *Proc. Irish Signals Systems Conf. (ISSC)*, 2012, pp. 1–5.

[6] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation." *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 71–82, 2013.

[7] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2010, pp. 246–253.

[8] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[9] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Inf. Process. Systems*, 2000, pp. 556–562.

[11] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation (article id 872425)," *Comput. Intell. Neurosc.*, vol. 2008, 2008.

[12] A. Cichocki, R. Zdunek, and A. H. Phan, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley and Sons2009.

[13] P. Smaragdis, "User guided audio selection from complex sound mixtures," in *Proc. ACM Symposium User Interface Software Technology (UIST)*, 2009, pp. 89–92.

[14] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Process. Magazine*, vol. 31, no. 3, pp. 116–124, May 2014.

[15] S. Abdallah and M. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2004, pp. 318–325.

[16] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, 2014.

[17] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. Int. Symp. Frontiers Res. Speech Music*, 2007, pp. 8–9.

[18] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, 2007.

[19] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, 2007.

[20] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, 2010.

[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoust., Speech Signal Process. (ICASSP), 2013 IEEE Int. Conf. on*. IEEE2013, pp. 7092–7096.

[22] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *ISMIR*, 2014, pp. 477–482.

[23] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoust., Speech Signal Process. (ICASSP), 2012 IEEE Int. Conf. on*. IEEE2012, pp. 57–60.

[24] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *Acoust., Speech Signal Process. (ICASSP), 2015 IEEE Int. Conf. on*. IEEE2015, pp. 718–722.

[25] D. Fano Yela, S. Ewert, D. FitzGerald, and M. B. Sandler, "Interference reduction in music recordings combining kernel additive modelling and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017.

[26] M. Müller and F. Kurth, "Enhancing similarity matrices for music audio analysis," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2006, pp. 437–440.

[27] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Int. Conf. Latent Variable Analysis Signal Separation*. Springer2017, pp. 323–332.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.