Drumloop Separation using adaptive Spectrogram Templates

Christian Dittmar[1], David Wagner[2], Daniel Gärtner[1]

[1] *Fraunhofer Institut für Digitale Medientechnologie, 98693 Ilmenau, Germany, Email: dmr,gtr@idmt.fraunhofer.de*

[2] *Technische Universität Ilmenau, 98693 Ilmenau, Germany, Email: david.wagner@tu-ilmenau.de*

Introduction

The separation of drumsounds from drumloops is a desirable signal processing functionality with a wide variety of applications in music production and music video games. Here, drumsounds describes the sound that is audible, when a drum or percussion instrument is hit. Since recognition of the involved drums is a prerequisite for separation, the detection of instrument types and corresponding onsets is necessary. Although machine-learning based classification of isolated drumsounds has been proven do be feasible, it is not directly applicable to the problem of drumloop separation. The main challenge is the strong overlap of drum spectra when two or more drumsounds share the same onset. It leads to erroneous estimation of the involved instruments, e.g., a tom and a hi-hat appearing simultaneously could easily be misclassified as being a snare. Different approaches have been proposed in the literature to overcome that problem, mainly template matching [1],[2] vs. decomposition based methods [3],[4],[5]. We pursue the approach of template matching, but combine it with a Non-Negative Matrix Factorization (NMF) in order to derive an initial estimate of the spectrogram templates of the involved drumsounds. A heuristic update rule for the templates is described as well as an expectation-maximization approach to the quasi-transcription. Due to space limits in this publication, a formal evaluation is omitted.

Drumloop separation

The proposed separation method exploits the fact that the most important drums have multiple onsets throughout a drumloop recording. Most of the time, they coincide with onsets of other drums. However, there may partly isolated occurences of the drumsound. By finding all onsets of that drum, it is possible to distill its magnitude spectrogram template and use the original phase spectrogram to resynthesize it into the time domain. As with melodic instruments, small sound variations between the different onsets are effectively captured in the phase spectrogram, thus allowing the usage of a quasi-static "mother" spectrogram.

Detection of onset candidates

The audio signal is transformed into a time-frequency representation using Short-Term Fourier Transform (STFT) with 46ms blocksize and 9ms hopsize. This yields the magnitude spectrogram X and the phase spectrogram Φ. Additionally, a spectral envelope representation V is computed by accumulating the energy in 25 critical bands covering the frequency range of X. Normalization of each frequency band by its standard deviation guarantees for equal importance of all drums. Onset candidates t are derived from X by means of peak picking in the relative difference function. A dynamic threshold is used to discard small maxima. The spectral envelopes V_t corresponding to the onset times t are stored for later processing.

Estimation of drum candidates

Decomposition of V into d_{max} components is conducted via NMF, as described in [3]. The NMF-model is given by $V \approx SA$, where S represents the basis spectra (spectral envelopes) of the involved drums and A represents their time-varying gains (amplitude envelopes). The reduced spectrogram V is used as input to NMF in order to spare computation time. An appropriately chosen subset of V_t is used to initialize S by sorting all V_t ascending by their spectral centroid and picking d_{max} spectra. This way, the iterative update of A and S gets an initial push into the desired direction. Consequently, the solution of the NMF is invariant (in contrast to random initialization) and there is no permutation problem. In each row of A (i.e., for the d-th drum candidate), the time-varying gains around t is accumulated and interpreted as onset probabilities $p_{t,d}$.

Spectrogram template adaption

Drum spectrogram template adaption was introduced in [1]. It is based on virtually stacking all spectrogram excerpts observed at one drum candidate's onsets on top of each other and taking the median for every time-frequency element. In this work, a similar distillation approach is pursued. For each drum candidate, a 500ms excerpt $X_{t,d}$ is taken from X at every onset time t. The very first excerpt serves as an initial estimate of the spectrogram template X_d and is chosen based on several heuristics (e.g., maximal $p_{t,d}$). The median calculation is replaced by the more efficient minimum operation. Of course, taking the minimum bears the danger of discarding too much information. Thus, every $X_{t,d}$ contributing to X_d is multiplied with a boost factor ensuring that the mass of the most important frequencies (usually a confined area around the frequency bin of the absolute maximum) is retained. Although a preliminary quasi-transcription has been derived by the preceding stages, it is still necessary to check carefully, which $X_{t,d}$ can be grouped together in order to distill the spectrogram template X_d. Therefore, Pearsons' correlation coefficient $r = corr(X_d, X_{d,t})$ is computed as a similarity measure. To account for slight variations in the drum onsets, r is computed multiple times for intentionally manipulated versions of $X_{d,t}$. The-

se manipulations comprise sub-frame time shift, sub-bin frequency shift, frequency stretching and nonlinear compression of the magnitude. The shift parameter combination that maximizes r is stored for reuse in the resynthesis stage. One important aspect of the spectrogram template adaption is the removal of subsequent drums occurring inside an excerpt. Based on the assumption that only drums with a sharp attack and fast decay occur, extrapolation of the decay slope overwrites the magnitude (which may contain the attack of the follow-up drums). The artificially continued decay slope of the drum is derived by means of linear regression. Beforehand, the magnitude of $X_{d,t}$ is logarithmized, thus a linear slope corresponds to an exponential decay in the original amplitude domain, which is a reasonable assumption for drums.

Quasi-transcription

This processing step shall derive a reliable quasi-transcription before starting with the resynthesis. It aims at deriving more realistic estimates of the onset probabilities $p_{t,d}$) without erroneous entries originating from crosstalk effects of other drums. Therefore, the X_d resp. their spectral envelope counterparts V_d are used. Intuitively, it becomes clear that 2D spectrogram templates are more robust for that task than single basis spectra of S, since the time-frequency progression of a drum shows more distinctive characteristics than its basis spectrum alone. The update of $p_{t,d}$ is computed by means of an expectation-maximization procedure. The assumption is, that V can be approximated as a weighted sum of the previously distilled V_d, placed at every onset point t. Thus, the expectation step consists of accumulating all V_d weighted with the initial $p_{t,d}$ at every t into $\hat{V}$. The model likelihood is given by $\tilde{V} = V./\hat{V}$, where ./ denotes element-wise division. In the maximization step, the contribution of each V_d to $\tilde{V}$ is used to update the $p_{t,d}$.

Optimal amplification and resynthesis

The resynthesis is realized by inverse STFT (overlap add method). For each probable onset, an optimally manipulated $\tilde{X}_d$ is derived from X_d by incorporating the previously stored shift parameters. Since the $\tilde{X}_d$ is arbitrarily scaled, it is necessary to estimate an optimal amplification factor before subtracting it locally from X. This factor is computed iteratively for each time frame based on a simple heuristic. It is assumed, that at the time of the onset, the original spectrogram X is made up from a mixture of $\tilde{X}_d$ and some other spectral components. By means of normalization, it is ensured that every frame of $\tilde{X}_d$ has equal mass to the corresponding frame of X. A constant boost factor guarantees, that it even surpasses the spectrum in X in certain frequency areas. The total mass of this overshoot is accumulated and subtracted from the total mass of the template spectrum in that frame. The ratio between overshoot-corrected and original mass gives a correction factor that is multiplied with that template spectrum. In practice, this algorithm converges quickly to a satisfactory amplification factor that adapts the template spectrum closely to the mixture spectrum.

Capabilities and limitations

The proposed method works well with input material of moderate complexity i.e., 3-5 different drums with consistent playing style. The method breaks down, when there is too much variation in the onsets of a certain drum (e.g., expressive brush playing on a snare drum). Informal listening tests revealed that idiophones like cymbals are only of limited separability. They usually exhibit broadband spectra that show a lot of variation and overlap strongly with other drums' spectra. Also, their extensive decay generates constant background noise that makes the distillation of the other drums difficult. In general, the usage of the original phase spectrogram for resynthesis may lead to audible artifacts.

Conclusions

A novel method for separation of drumsounds from drumloop recordings has been presented. Informal listening tests show, that it provides unprecedented separation audio quality for drumloops of moderate complexity. Future work will be directed towards using alternative time-frequency transformations and finding more robust update rules for the drum spectrogram templates. In principle, the presented method is applicable for quasi-transcription of drums from polyphonic music as well. It is up to further investigation, if the assumptions made for drumloops can be transferred to real-world music recordings.

Acknowledgments

The authors would like to thank the team of Celemony Software GmbH for fruitful discussions and valuable input to this work.

Literatur

[1] Yoshii K., Goto M., Okuno H.: Automatic Drum Sound Description for Real-World Music Using Template Adaptation and Matching Methods. Proceedings of ISMIR, Barcelona, Spain, Oct. 10-14. 2004.

[2] Gillet O., Richard G.: Transcription and Separation of Drum Signals From Polyphonic Music. IEEE Transactions on Audio, Speech, and Language Processing, 529-540, 2008.

[3] Helen M., Virtanen T.: Separation of Drums From Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine. Proceedings of EUSIPCO, Antalya, Turkey, Sept. 4-8. 2005.

[4] Dittmar C., Uhle, C.: Further Steps towards Drum Transcription of Polyphonic Music. Proceedings of 116th International AES Convention, Berlin, Germany, May 8-11. 2004.

[5] Fitzgerald D., Coyle E., Cranitch M.: Using Tensor Factorisation Models to Separate Drums from Polyphonic Music. Proceedings of the International Conference on Digital Audio Effects, Como, Italy, Sept. 1-4. 2009.