Generalized Cepstral Features for Clustering in Blind Audio Source Separation

Christian Rohlfing[1], Julian Mathias Becker[1]

[1] *Institut für Nachrichtentechnik, RWTH Aachen University, 52056 Aachen, Email: rohlfing@ient.rwth-aachen.de*

Abstract

To generalize cepstral domain audio features, the usage of the so-called generalized logarithm function has been proposed. In this paper, the A-law companding function is suggested as another suitable generalization regarding amplitude scaling and mel-scale warping. The application of these generalized cepstral features in a state-of-the-art blind source separation (BSS) algorithm is evaluated: The non-negative matrix factorization (NMF) is both used for separating the mixture into single note events and for clustering these events into the estimations of the single sources. While achieving comparable separation quality, the proposed A-law methods do not need an extra preprocessing step to ensure non-negativity of the feature values.

Note Separation with NMF

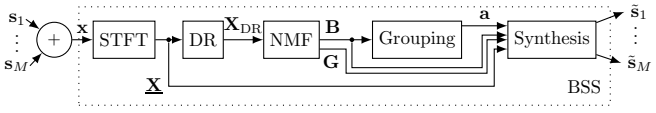

Figure 1: Flowgraph of the BSS algorithm.

Fig. 1 shows the flow graph of the used BSS algorithm which is described in detail in [1]:
The mixture $\mathbf{x} = \sum_{m=1}^{M} \mathbf{s}_m$ is a linear sum of independent monaural sources $\mathbf{s}_m$ in the time domain. After applying the short-time Fourier transform (STFT), a spectral dimension reduction (DR) of the mixture amplitude spectrogram $|\underline{\mathbf{X}}|$ is performed using a mel-filter bank.
Afterwards, the dimension-reduced spectrogram $\mathbf{X}_{\mathrm{DR}}$ is factorized by the NMF into I acoustical events (notes) by approximating $\mathbf{X}_{\mathrm{DR}} \in \mathbb{R}_+^{K \times T}$ as a matrix product $\mathbf{X}_{\mathrm{DR}} \approx \mathbf{BG}$. $\mathbf{B} \in \mathbb{R}_+^{K \times I}$ contains the frequency basis vectors of each event whereas $\mathbf{G} \in \mathbb{R}_+^{I \times T}$ consists of the corresponding time envelopes [1]. K denotes the number of reduced frequency bins and T the number of time bins. $\mathbf{B}$ and $\mathbf{G}$ are initialized following a semantic initialization and are estimated using the β-NMF [1].

Feature Calculation

In [4], the generalized cepstrum is proposed. The logarithm, which is typically applied for the amplitude scaling, is replaced by the Box-Cox-Transformation (BCT), which generalizes the natural logarithm

$$F_{\mathrm{BCT}}(x, \lambda) = \begin{cases} \ln x & \text{if } \lambda = 0\,, \\ \left(x^{\lambda} - 1\right)/\lambda & \text{otherwise.} \end{cases} \tag{1}$$

In [1], this approach is used to obtain generalized mel-frequency cepstral coefficients (MFCC) for clustering the acoustical events to the estimated sources:

First, each column of $\mathbf{B}$ is frequency-warped by filtering with a generalized mel-filter bank which consists of K_{mel} overlapping triangular-shaped filters. These filters are spaced corresponding to a generalized mel-scale [1]:

$$f_{\mathrm{mel}} = F_{\mathrm{BCT}}\left(f_{\mathrm{Hertz}}/700 + 1, \lambda_f\right)\,, \tag{2}$$

with f_{Hertz} corresponding to the linear frequency scale and f_{mel} to the warped mel scale.
Afterwards, the resulting coefficients are stored into the feature matrix $\mathbf{F}(k_{\mathrm{mel}}, i)$ with k_{mel} being the filter bank index and i indexing the acoustical events. The feature matrix is normalized to its maximum amplitude

$$\mathbf{F}(k_{\mathrm{mel}}, i) \leftarrow \mathbf{F}(k_{\mathrm{mel}}, i) / \max_{k_{\mathrm{mel}}, i} \mathbf{F}(k_{\mathrm{mel}}, i) \tag{3}$$

and then scaled by applying the BCT function again [1]:

$$\mathbf{F}(k_{\mathrm{mel}}, i) \leftarrow F_{\mathrm{BCT}}\left(C\mathbf{F}(k_{\mathrm{mel}}, i) + 1, \lambda_a\right)\,. \tag{4}$$

The $+1$ in the argument of the BCT in Eq. (4) ensures non-negativity of the values which is necessary for clustering with the NMF. Also, small positive values get mapped onto small positive values instead of large negative values when omitting the $+1$. The multiplication with the factor C is necessary to revert the influence of the $+1$ on the shape of the logarithm [1].
The decorrelation step, which is typically done during the calculation of MFCCs by applying the discrete cosine transform, is omitted, as it does not increase the separation quality [1].

Clustering and Synthesis

The NMF is used again to cluster the feature matrix $\mathbf{F} \in \mathbb{R}^{K_{\mathrm{mel}} \times I}$: $\mathbf{F} \approx \mathbf{CH}$. $\mathbf{C} \in \mathbb{R}^{K_{\mathrm{mel}} \times M}$ contains the M clusters and $\mathbf{H} \in \mathbb{R}^{M \times I}$ the corresponding soft cluster assignments. By calculating $\mathbf{a}(i) = \arg\max_m \mathbf{H}(m, i)$, a hard cluster decision is obtained which is used during the synthesis step. An SVD-based initialization for $\mathbf{C}$ and $\mathbf{H}$ is used as described in [1].
For synthesis, the complex spectrogram $\underline{\mathbf{Y}}_i$ of each separated event has to be obtained. The scheme described in [1] is used here. The estimated sources can then be calculated as $\tilde{\mathbf{s}}_m = \sum_{\mathbf{a}(i)=m} \mathbf{y}_i$, where $\mathbf{y}_i$ denotes the time-domain signal of each event and is calculated as the inverse STFT of $\underline{\mathbf{Y}}_i$.

Feature Calculation with A-Law Function

In this paper, the usage of the A-law companding algorithm [2] for amplitude scaling is proposed to get rid of the normalization parameter C in Eq. (4). A similar procedure was introduced in [3] for speech recognition which uses the μ-law algorithm instead.

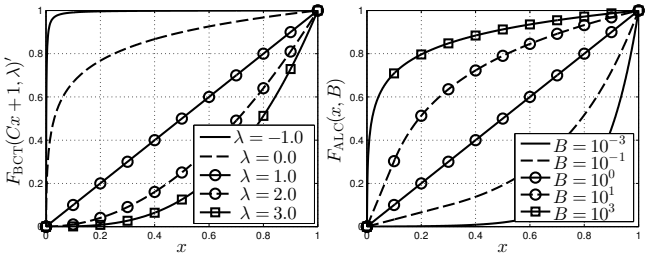

Figure 2: BCT with non-negativity constraint on the left and A-law companding curves on the right. Each BCT curve has to be normalized by its maximal value $F_{\mathrm{BCT}}(C+1,\lambda)$ to fit in the range $[0,1]$.

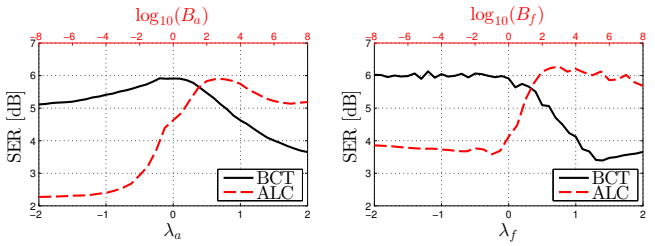

Figure 3: Signal-to-error ratio (SER) for amplitude scaling with BCT and ALC with $\lambda_f = 0$ on the left and for frequency warping with $\lambda_a = 0$ on the right.

The convex A-law companding function is defined as

$$F_{\mathrm{AL}}(x,A) = \frac{1}{1+\ln A}\begin{cases} Ax & \text{if } 0 \le x < \frac{1}{A}\,, \\ 1+\ln(Ax) & \text{if } \frac{1}{A} \le x \le 1\,, \end{cases} \tag{5}$$

The inverse concave expanding function is given as

$$F^{-1}_{\mathrm{AL}}(y,A) = \frac{1}{A}\begin{cases} (1+\ln A)y & \text{if } 0 \le y < \frac{1}{1+\ln A}\,, \\ \mathrm{e}^{y(1+\ln A)-1} & \text{if } \frac{1}{1+\ln A} \le y \le 1\,, \end{cases} \tag{6}$$

with $A \ge 1$. Eq. (5) and (6) are piecewise linear. To combine convex and concave behaviour similar to the BCT, a combined A-law function (ALC) is proposed:

$$F_{\mathrm{ALC}}(x,B) = \begin{cases} F_{\mathrm{AL}}\left(x,B\right) & \text{if } B \ge 1\,, \\ F^{-1}_{\mathrm{AL}}\left(x,\frac{1}{B}\right) & \text{if } 0 < B < 1\,. \end{cases} \tag{7}$$

Fig. 2 shows BCT curves with the non-negativity constraint (cf. Eq. (4)) with $C = 999$ on the left for different values of λ and ALC curves (cf. Eq (7)) with different values for B on the right.
Using F_{ALC} for amplitude scaling, Eq. (4) transforms to

$$\mathbf{F}(k_{\mathrm{mel}},i) \leftarrow F_{\mathrm{ALC}}\left(\mathbf{F}(k_{\mathrm{mel}},i),B_a\right)\,, \tag{8}$$

which needs only one parameter (B_a) compared to Eq. (4) which needs two parameters (λ_a, C).
It is also possible to use Eq. (7) for frequency warping:

$$f_{\mathrm{mel}} = F_{\mathrm{ALC}}\left(f_{\mathrm{Hertz}}/f_{\max},B_f\right)\,, \tag{9}$$

with $f_{\max}$ denoting the maximum frequency.

Evaluation

For evaluation, a test set which is described in [1] is used. It consists of 60 monaural recordings of different instruments, vocals, speech and noise. Only mixtures with two different sources are considered ($M = 2$), which results in a total of 1770 different mixtures.
Taking possible dynamic differences between the sources into account, three different runs of the algorithm are conducted for each mixture with signal power differences of 0 dB and ±12 dB.
The STFT and ISTFT are used with a window size of 4096 samples and an overlap of 50 %. The dimension reduction results in a frequency resolution of $K = 400$.

The NMF is applied for note separation with $I = 20$ and $\beta = 0$ and stops after 300 iterations. The clustering NMF operates with $\beta = 1$ and stops after 100 iterations.
$K_{\mathrm{mel}} = 20$ mel-filter banks are used for frequency warping. The normalization parameter in Eq. (4) is set to $C = 999$ which sets the range of $\mathbf{F}(k_{\mathrm{mel}},i)$ to 60 dB [1].
For all simulations, the BCT and the A-law parameter are chosen as $\lambda_{f/a} \in [-2,2]$ and $B_{f/a} \in [10^{-8},10^{8}]$.
Fig. 3 shows separation results regarding the amplitude scaling on the left and frequency warping on the right. For the amplitude scaling, the ALC gives roughly the same maximal SER as the BCT. The maxima are reached for $\lambda_a \approx 0$ and $B_a \approx 10^3$ respectively. The SER of the ALC regarding the frequency warping is even increased compared to the BCT (about 0.15 dB) which can be explained by the linear region of Eq. (7).
Using the BCT or the ALC for both amplitude scaling and frequency warping, the maximum SER of 6.09 dB for the BCT is achieved for $(\lambda_a,\lambda_f) = (-0.5,-0.5)$ and of 6.26 dB for the ALC for $(B_a,B_f) = (10^3,10^3)$.

Conclusion

A modification of the generalized mel-frequency cepstral coefficients was proposed: The A-law algorithm instead of the Box-Cox-Transformation is used for both amplitude scaling and frequency warping. These novel features are evaluated in the scope of blind source separation algorithm: The novel approach needs one parameter less and obtains slightly better separation results due to a linear behaviour for small amplitudes.

References

[1] Spiertz, M.: Underdetermined Blind Source Separation for Audio Signals. Aachen Series on Multimedia and Communications Engineering, Shaker Verlag, Aachen, 2012.

[2] Gersho, A.: Principles of quantization. IEEE Transactions on Circuits and Systems 25 (1978), 427-436.

[3] Haderlein, T., Stemmer, G., Nöth, E.: Speech recognition with μ-law companded features on reverberated signals. Text, Speech and Dialogue. Springer Berlin Heidelberg, 2003, 173-180.

[4] Kobayashi, T., Imai, S.: Spectral analysis using generalized cepstrum. IEEE Transactions on Acoustics, Speech and Signal Processing 32.5, 1984, 1087-1089.