Practical work on audio source separation

Roland Badeau

In this practical work, you will program a simple Matlab implementation of a variant of the DUET (*Degenerate Unmixing Estimation Technique*) method, which aims to separate sound sources in a stereophonic mixture. You will thus address the case of an under-determined ($M = 2$ sensors and $K > 2$ sources) instantaneous linear mixture. The separation is achieved by exploiting the spatial information, and by assuming that the sources are sparse in the time-frequency plane (a single source is active at each time-frequency bin). The transformation that you will use is the MDCT (*Modified Discrete Cosine Transform*), which presents the double advantage of having real values, and of producing a sparser representation than the STFT (short time Fourier transform). In order to compute it, you will use the *Linear Time/Frequency Toolbox* (`ltfat`) toolbox, that you can download on the website of M2 MVA. You will also find on this website the stereophonic sound file to be processed, named `mix.wav`. In order to use the functions of the `ltfat` toolbox, you will first have to load it by calling function `ltfatstart`.

1 Mixture model and principle of the DUET method

The DUET method relies on the following mixture model: at every time-frequency bin (f, n),

$$\boldsymbol{X}(f,n) = \boldsymbol{S}(f,n)\,\boldsymbol{A}$$

where

- the row vectors $\boldsymbol{X}(f,n) = [X(f,n,1), X(f,n,2)]$ of dimension $M = 2$ contain the MDCT of the two stereophonic channels $x(t,m)$ of the observed mixture;
- the $K \times M$ matrix $\boldsymbol{A} = [\cos(\boldsymbol{\theta}), \sin(\boldsymbol{\theta})]$ is the mixing matrix;
- the K-dimensional column vector $\boldsymbol{\theta}$ contains the angles $\theta(k)$ of sources k;
- the K-dimensional row vectors $\boldsymbol{S}(f,n) = [S(f,n,1), \ldots, S(f,n,K)]$ contain the MDCT of the K unknown source signals.

If only source k is active at (f,n), the point of affix $Z(f,n) = X(f,n,1) + iX(f,n,2) \in \mathbb{C}$ is such that $Z(f,n) = S(f,n,k)\,e^{i\theta(k)}$, where $S(f,n,k) \in \mathbb{R}$. We remark that its argument permits us both to identify the active source k at (f,n) and its angle $\theta(k)$. The magnitude of $Z(f,n)$ permits us to determine the value of $S(f,n,k)$, up to its sign. In order to remove the sign ambiguity of $S(f,n,k)$, we can assume that $\theta(k) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Once source k is identified at every time-frequency bin, a binary mask $B(f,n,k)$ can be applied to $X(f,n,m)$, in order to obtain an estimation $Y(f,n,m,k)$ of the stereophonic image of source k. The source signal is finally reconstructed by means of the MMSE (*Minimum Mean Square Error*) estimator, which is such that $S(f,n,k) = Y(f,n,1,k)\cos(\theta(k)) + Y(f,n,2,k)\sin(\theta(k))$.

2 Work to do

1. Open file `mix.wav` and load it in a $T \times M$ matrix $x(t,m)$, where $M = 2$ and T is the number of samples. Use your headphones to listen to the mixture. What is the number K of instruments that you can hear? From which direction do you perceive them?
2. Plot the temporal dispersion diagram, defined as the set of points in the plane of coordinates $(x(t,1), x(t,2))$ for all t (in order to plot a set of points, you can use function `plot` with parameter `'x'`, and you can normalize the axes with the instruction `axis equal`). Can you distinguish the directions of the sources?

3. Compute the MDCT $X(f,n,m)$ of the two stereophonic channels $x(t,m)$ (you can use function `wmdct`, with $F = 512$ frequency bands, and the window `'sqrthann'`). Plot the corresponding time-frequency representations $|X(f,n,m)|^2$ (you can use function `plotwmdct`).

4. Plot the time-frequency dispersion diagram, defined as the set of points in the plane of affix $Z(f,n)$ for all f and n. Can you distinguish the directions of the sources? How do you explain it?

5. Plot the histogram of the arguments of the points of affix $Z(f,n)$ for all f and n (you can use function `atan` to compute the arguments modulo π, between $-\frac{\pi}{2}$ and $+\frac{\pi}{2}$, and function `hist` to compute the histogram, whose number of classes has to be tuned so as to make the directions of the sources clearly visible). Estimate the angles $\theta(k)$ (you can determine these values graphically from the histogram).

6. In order to estimate the active source at every time-frequency bin (f,n), you can look for the source k whose angle $\theta(k)$ is closest to the argument of $Z(f,n)$, modulo π (you can use a deviation measure invariant modulo π, for instance $|\sin(\theta(k) - \angle Z(f,n))|$). Then generate the binary masks $B \in \{0,1\}$, such that $B(f,n,k)$ is equal to 1 if source k is active at (f,n), or 0 otherwise.

7. Apply masks B to the MDCT $X(f,n,m)$ in order to estimate the MDCT of the stereophonic images $Y(f,n,m,k)$. Then reconstruct the images $y(t,m,k)$ of the source signals by applying the inverse MDCT (you can use function `iwmdct`).

8. Listen to the K reconstructed stereophonic images $y(:,:,k)$. What defects can you perceive?

9. Compute the MMSE estimator $S(f,n,k)$ of source k. Reconstruct the source signals $s(t,k)$ by applying the inverse MDCT to $S(f,n,k)$. Listen to the result.

10. We now wish to respatialize the sources, i.e. to resynthesize the mixture $x(t,m)$ by modifying the angles $\theta(k)$ (remark that it is not needed to switch back to the MDCT domain). For instance, try to permute the directions of the sources. Listen to the result. What audible defects can you notice?