Audio Source Separation by the DUET Method

Practicum report

TSIA 206 — Speech and Audio Processing · Télécom Paris

Hussein El Gouch & **Lukas Tabouri**

1 Objective and design

The goal of this session was to separate K sound sources from an *under-determined* ($M = 2$ sensors, $K > 2$ sources) instantaneous stereophonic mixture, using only the spatial information and the assumption that the sources are sparse in the time–frequency (TF) plane. We implemented a variant of **DUET** (*Degenerate Unmixing Estimation Technique*).

Our design follows the DUET principle directly:

- We transform both channels with the **MDCT** rather than the STFT, because it is real-valued and yields a sparser representation, which makes the *W-disjoint orthogonality* assumption (one dominant source per TF bin) more realistic.
- We form the complex affix $Z(f, n) = X(f, n, 1) + i\, X(f, n, 2)$. Where a single source k is active, $Z(f, n) = S(f, n, k)\, e^{i\theta(k)}$: its *argument* encodes the source direction $\theta(k)$ and its *magnitude* encodes the amplitude.
- The source angles are estimated from a magnitude-weighted histogram of $\arg Z \pmod{\pi}$. Each TF bin is then assigned to the nearest source with a binary mask, using the π-invariant deviation $|\sin(\theta(k) - \angle Z(f, n))|$.
- Stereophonic images are obtained by masking + inverse MDCT; the mono sources are recovered with the MMSE projector $S(f, n, k) = Y(f, n, 1, k) \cos\theta(k) + Y(f, n, 2, k) \sin\theta(k)$; finally the mixture is respatialised *directly in the time domain*.

2 Implementation

The pipeline was implemented in Python (the `mdct` toolkit on top of `scipy`/`numpy`). The mixture `mix.wav` is stereo, 22.05 kHz, $T \approx 2.2 \times 10^5$ samples; the MDCT uses $F = 512$ bands, giving an X tensor of shape (256, 863, 2).

The key implementation choice was the **analysis/synthesis window**. We used the sine (`cosine`) window, i.e. the perfect-reconstruction window satisfying the Princen–Bradley condition $w[n]^2 + w[n + N]^2 = 1$ — the Python equivalent of LTFAT's `'sqrthann'`. With it the analysis–synthesis round-trip error is $\sim 10^{-14}$. Masks are built by broadcasting the deviation over (f, n, k) and taking the `argmin` along k, which is fully vectorised.

The analysis identified $K = 3$ **sources** at $\theta \approx [-75°, -15°, +45°]$, i.e. one panned on each channel and one in the centre. The diagnostics confirmed the model: the *temporal* dispersion diagram is a diffuse cloud (no readable directions), whereas the *time–frequency* dispersion of Z collapses onto three clear rays at the estimated angles. Images, MMSE sources and the respatialised mixture were all reconstructed and auditioned.

3 Encountered difficulties

- **Window / perfect reconstruction.** The starting template used a `hamming` window, which does *not* satisfy Princen–Bradley: it gave a $\approx 25\%$ round-trip error and audible artefacts *before* any masking. Identifying this and switching to the sine window was the main subtlety, and it materially improved the separation quality.
- **Legacy toolchain.** `mdct 0.4` and its `stft` backend were written for an old `scipy`/`numpy` (re-

moved symbols such as `scipy.signal.kaiser` and `scipy.real`, and list-based array indexing). We wrote a small compatibility shim so the notebook runs on current versions.

- **Choosing K.** The histogram shows three sharp peaks but also a faint shoulder near $-23°$. Deciding that this was the skirt of the dominant $-15°$ source — rather than a fourth instrument — required comparing peak energies and listening to the candidate separations.
- **Histogram tuning.** The number of bins and the magnitude weighting had to be tuned to make the three directions clearly visible without fragmenting a single peak into several.

4 Reflection: where the real challenges lie

The mechanical pipeline is short; the genuine difficulty is conceptual and lives in the *assumptions*, not the code.

- **W-disjoint orthogonality is only approximate.** Real instruments share harmonics, so many TF bins mix several sources. Every such bin is mis-assigned by the hard mask, which is what produces the characteristic *musical noise* ("birdies"), inter-source leakage, and spectral holes. This — not the implementation — is the dominant limit on quality.
- **Binary masking is a blunt estimator.** A soft / Wiener-type mask would trade musical noise for some additional leakage; the binary choice maximises simplicity at the price of artefacts.
- **Angle estimation is fragile.** Directions are read from a histogram; close sources, weak sources, or reverberation would blur the peaks. Our mixture is anechoic and well spread in angle, which is the favourable case — a realistic recording would be much harder.
- **The geometry is under-determined.** With $M = 2$ sensors and $K = 3$ sources there is no algebraic inverse; separation is *only* possible because of sparsity, so everything rests on that statistical assumption holding.
- **Evaluation is subjective.** No reference sources are provided, so quality is judged by ear; objective SDR/SIR/SAR metrics would require ground-truth signals.

Respatialisation makes these limits audible: once the sources are moved, the leaked and mis-assigned energy is placed at the wrong location ("ghost" sources), and the component discarded by the MMSE projection (about 16% of the mixture energy in our run) is missing — so the resynthesised mix is not identical to the original. In short, DUET is an elegant, minimal answer to a hard problem, and its performance is governed almost entirely by how well the sparsity assumption holds. Maximising that disjointness — through the choice of transform and of mask — is where the real engineering challenge of source separation sits.