Speech-guided source separation using a pitch-adaptive guide signal model

Romain HENNEQUIN, Juan-José BURRED, Simon MALLER, Pierre LEVEAU

Audionamix, 171 quai de Valmy, 75010 Paris, France

Introduction

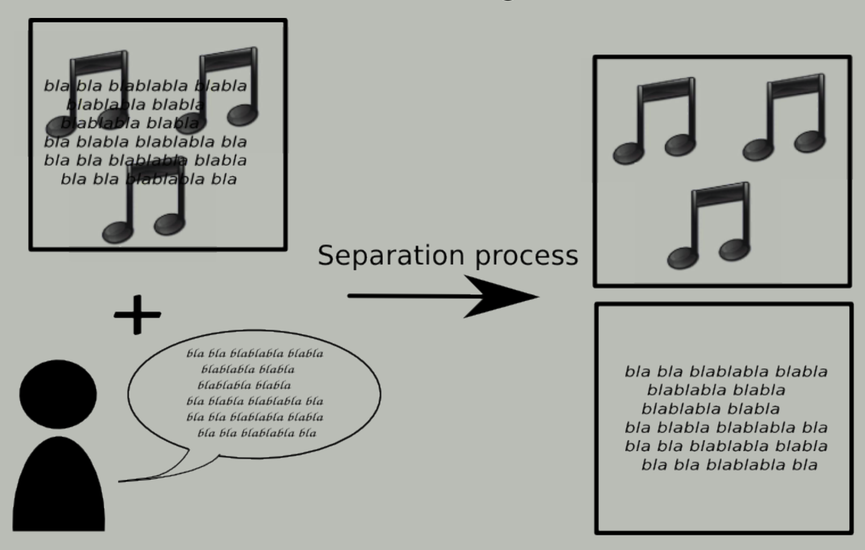
New method to perform underdetermined audio source separation using a spoken or sung reference signal to inform the separation process.

- Explicitly models differences between reference and target signal.
- Outperforms state-of-the-art methods.

Typical scenario

Isolate a target sound, imitated by the user, from a background sound.

- Isolation/removal of dialogs in a movie soundtrack.
- Isolation/removal of lead vocal in a song.



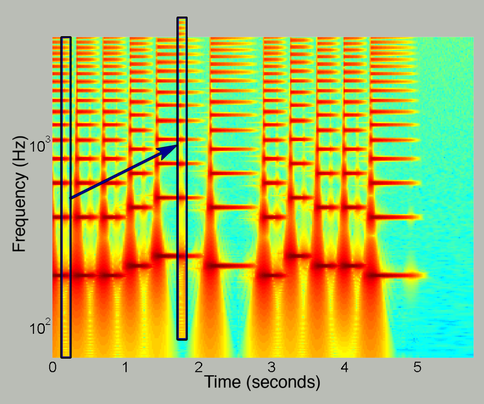
Issues: differences between guide and target signal:

- pitch differences: absolute pitch, intonation, vibrato...
- time differences: misalignments.
- equalization/timbre differences.

Model

Input spectrogram

$\mathbf{V}$: power log-frequency spectrogram (Constant-Q transform) of the mixture signal.
Shift invariance property: a pitch modification can be modeled by a vertical shift.



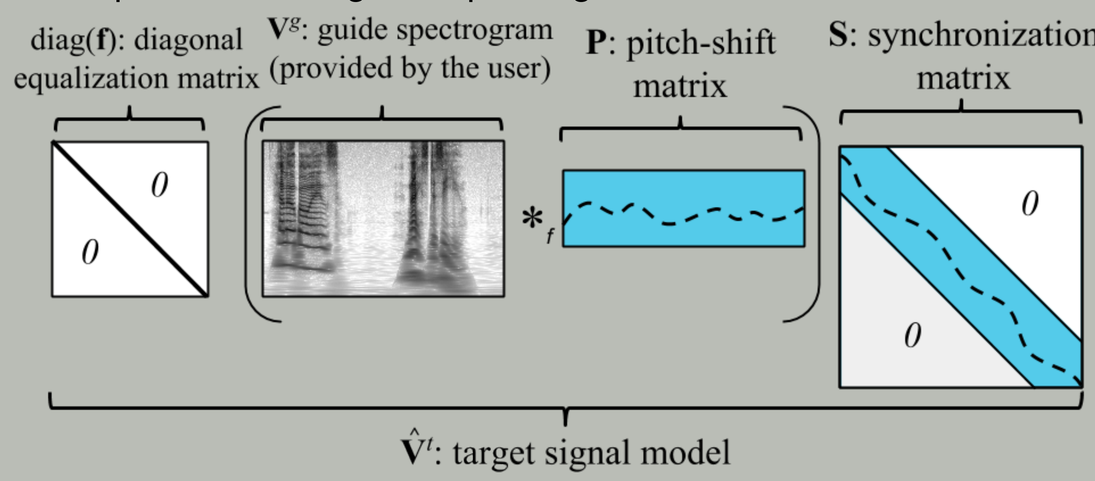
Non-negative spectrogram model

$$\mathbf{V} \approx \hat{\mathbf{V}} = \hat{\mathbf{V}}^t + \hat{\mathbf{V}}^b$$

$\hat{\mathbf{V}}^t$: target signal model.
$\hat{\mathbf{V}}^b$: background signal model.

Target signal model:

$\hat{\mathbf{V}}^t$ is adapted from the guide spectrogram $\mathbf{V}^g$:



Pitch shift operator $\mathbf{P}$:

$$\mathbf{V}^g_{\text{shifted}} = \left(\sum_{\phi} \overset{\downarrow\phi}{\mathbf{V}^g} \text{diag}(\mathbf{P}_{\phi,:}) \right)$$

Global adaptation filter $\mathbf{f}$ and synchronization matrix $\boldsymbol{S}$:

$$\hat{\mathbf{V}}^t = \text{diag}\,(\mathbf{f}) \left(\sum_{\phi} \overset{\downarrow\phi}{\mathbf{V}^g} \text{diag}(\mathbf{P}_{\phi,:}) \right) \mathbf{S}$$

Background signal model

Standard Non negative Matrix Factorization (generic model):

$$\hat{\mathbf{V}}^b = \mathbf{WH}$$

Algorithm

First estimation of parameters

Minimization of Itakura-Saito divergence:

$$\mathcal{C}(\mathbf{P}, \mathbf{f}, \mathbf{S}, \mathbf{W}, \mathbf{H}) = D_{IS}(\mathbf{V}|\hat{\mathbf{V}}^t + \hat{\mathbf{V}}^b), \text{ (s.t. } \mathbf{P}, \mathbf{f}, \mathbf{S}, \mathbf{W}, \mathbf{H} \geq \mathbf{0})$$

using multiplicative update rules.

Tracking

- Tracking of the best pitch shift in $\mathbf{P}$ using Viterbi algorithm.
- Tracking of the best synchronization in $\mathbf{S}$ using DTW.

Thresholding of $\mathbf{P}$ and $\mathbf{S}$ then parameters reestimation.

Separation

Separation is achieved using Wiener Filtering and CQT inversion.

Experimental results

Experimental setup

Dialog isolation/removal in movie soundtracks.
Compared with other separation algorithms:

- Main melody extraction method proposed in [1].
- PLCA-based speech-informed separation algorithm [2].
- First one but informed with instantaneous pitch of the target signal [3].
- CQT Wiener oracle as an upper performance bound.

Database

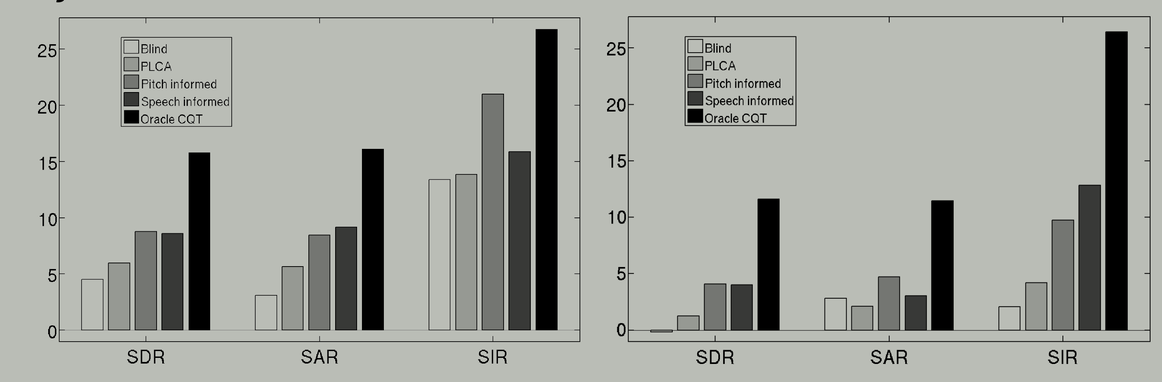
Mix signals:

- **10** excerpts from synthetic movie soundtracks.
- **5** different movies in english.
- Excerpt = dialog only part + music and effects only part mixed down to mono.

Dubs:

- using the mix signal as a reference.
- Done by the same male native english speaker.
- Same dubs for both speech informed algorithms.

Objective results

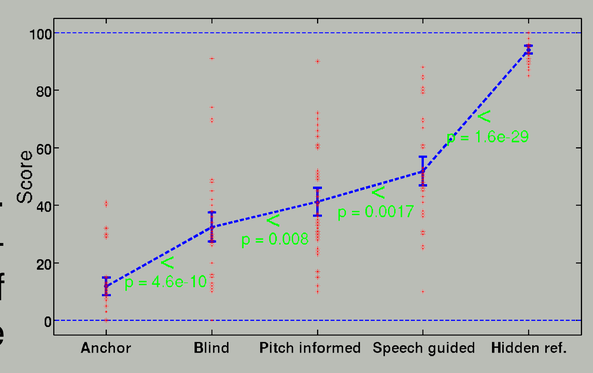


Listening tests

Internal blind listening test:

- Dialog isolation only.
- **5** (few!) participants.
- MUSHRA protocol.
- Rating according to the ”usability”.

Results of our algorithm globally preferred over the results of the pitch-informed algorithm (to be taken with care)



Conclusion

- New method to perform source separation providing a spoken guide.
- Outperforms a state-of-the-art one and methods performing same task.
- Future work:
  - No voice model on the guide signal: any kind of signal can be used. Many other applications.
  - Other kinds of adaptation (formant adaptation...).
  - Speeding up the algorithm.

References

J.-L. D. et al., “An iterative approach to monaural musical mixture de-soloing,” in *ICASSP*, 2009.

P. Smaragdis and G. J. Mysore, “Separation by humming: User-guided sound extraction from monophonic mixtures,” in *WASPAA*, 2009.

J.-L. Durrieu and J.-P. Thiran, “Musical audio source separation based on user-selected f0 track,” in *LVA/ICA*, 2012.