

049035 - Digital Speech Processing in Noisy Environments

Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria

Tuomas Virtanen

presented by Alexander Berkovich

Outline

- Definition of the problem
- Existing solution
- Article solution and results
- Conclusions
- Pro's and Con's
- Future work

Definition of the problem

We want to be able to “hear out” a single source as humans can.

$$X = \sum_i b_i$$

The problem – separating the sources b_i while having a single input X .

Existing solutions

1. Extract the most prominent source – F0

In article [2] an algorithm, “PreFEst“, finds the most predominant F0 frequency. The MAP estimation is used with the EM algorithm to estimate the PDF of the F0s. The novelty is that there is no assumption regarding the number of sources.

2. Segregate resolved and unresolved harmonics differently

In article [3], an algorithm is introduced that deals differently with resolved and unresolved harmonics (based on inspection of humans).

Existing solutions

3. Blind Source Separation using a set of known functions in the time domain.

In article [4] an algorithm is introduced that is using ML approach and a learned set of basic functions in the time domain.

Existing solutions

4. ICA - Finding the decomposition that reduced the statistical independence of the components or makes them nonredundant.

In article [5] an algorithm is introduced that applies

- PCA (ICA)
- Grouping some sources using a High Order Statistical Distortion
- Refining the HOSD clusters the components maximally similar or distant.

In article [6] it was shown that ICA algorithm, forcing independence up to 4th order statistics it better than the usual PCA algorithm

Existing solutions

5. Non negative Matrix Factorization

In article [7] an algorithm was introduced that decomposes the input X as $X=WH$, $W, H \in R^+$. It was shown that if X is the musical spectra, the rows of H are the temporal information and the columns of W are the spectrum information. (similar to ICA technique).

Existing solutions

6. Sparse coding

In article [8] new methods were introduced to extract the notes and the characteristics of notes, for transcription methods.

- Independent notes.
- Sparseness of the notes – a few are present at each time.
- ICA was applied.

In article [9] an algorithm for source separation was introduced that assumes:

- Sparseness of sounds (inactivity most of the time).
- Non-negativity (due to power spectra using).
- Temporal continuity between frames.
- The spectra of the sources is constant in time.

Article solution - The problem

$$x_t = \sum_{j=1}^J g_{j,t} b_j$$

X_t – power spectrum in frame t

B_j – basis function j – constant in time.

$g_{j,t}$ – gain of j -th basis function in frame t .

T – frame; J – number of basis function;

$$X = BG$$

$$X = [x_1 \dots x_T], B = [b_1 \dots b_J], G_{j,t} = g_{j,t}$$

Article solution – The Method

- **ICA (Independent Component Analysis)**— can't be applied directly due to lack of input.
- **ISA (Independent Subset Analysis)** — for each time frame t , the spectrum is seen as invariant of the phase and considered as the input.
- **NMF** — all spectra is nonnegative, and all gains are non-negative as well.
- **Sparseness** — the probability of g to be 0 is high.
- **Temporal Continuity**

Article solution

Minimizing the cost function:

$$c(B, G) = c_r(B, G) + \alpha c_t(G) + \beta c_s(G)$$

$c_r(B, G)$ - Reconstruction Error

$c_t(G)$ - Temporal continuity

$c_s(G)$ - Sparseness

α, β - weights

Article solution - Minimizing the cost function

Reconstruction: $c_r(B, G)$

$$c_r(B, G) = \sum_{k,t} X_{k,t} \log \frac{X_{k,t}}{[BG]_{k,t}} - X_{k,t} + [BG]_{k,t}$$

- Maximum Likelihood Estimator, when the observation is $\sim Po([BG]_{k,t})$
- Sensitive to low energy.
- Better for human observers.
- Linear in the input.

Article solution - Minimizing the cost function

Temporal continuity: $c_t(G)$

$$c_t(G) = \sum_{j=1}^J \frac{1}{\sigma_j^2} \sum_{t=2}^T (g_{t,j} - g_{t-1,j})^2$$

- Each frame isn't an individual observation.
- Normalization using

$$\sigma_j = \sqrt{\frac{1}{T} \sum_{t=1}^T g_{t,j}^2}$$

- The use of $(.)^2$ is due to better performance – maybe caused by the dependence of $\nabla[(.)^2]$ on the value and not only on the sign.

Article solution - Minimizing the cost function

Sparseness: $c_s(G)$

$$c_s(G) = \sum_{j=1, t=1}^{J, T} f(g_{j,t} / \sigma_j)$$

- Effective when a single model is used.
- $f(x)$ – punish nonzero g .
- $f(x) = \log(x^2+1)$, $f(x) = -\exp(-x^2)$, $f(x) = |x|$.
- $f(x) = |x|$. Less sensitive for the weight of this error term.

Article solution - Estimation

Update B:

$$B \leftarrow B \times \frac{\frac{X}{BG} G^T}{1G^T}$$

In [11] a proof that with this rule, C_r is non-increasing.

Update G using: $\nabla c(B, G)$

Article solution - Estimation – cont.

$$\nabla c(B, G) = \underbrace{B^T \mathbf{1}}_{\nabla c_r^+} - \underbrace{B^T \frac{X}{BG}}_{\nabla c_r^-}$$

$$+ \alpha \left[\underbrace{+\frac{4Tg_{j,t}}{\sum_{i=1}^T g_{j,i}^2}}_{\nabla c_t^+} - \underbrace{\left(2T \frac{g_{j,t-1} + g_{j,t+1}}{\sum_{i=1}^T g_{j,i}^2} + \frac{2Tg_{j,t} \sum_{i=2}^T (g_{j,i} + g_{j,i-1})^2}{\left(\sum_{i=1}^T g_{j,i}^2 \right)^2} \right)}_{\nabla c_t^-} \right]$$

$$+ \beta \left[\underbrace{+\frac{1}{\sqrt{\frac{1}{T} \sum_{i=1}^T g_{j,i}^2}}}_{\nabla c_s^+} - \underbrace{\frac{g_{j,t} \sqrt{T} \sum_{i=1}^T g_{j,i}}{\left(\sum_{i=1}^T g_{j,i}^2 \right)^{\frac{3}{2}}}}_{\nabla c_s^-} \right]$$

Article solution - Estimation – cont.

Update G :

$$G \leftarrow G \cdot \frac{\nabla c^-(B, G)}{\nabla c^+(B, G)}$$

This rule doesn't necessarily decrease the cost function

Article solution - Estimation – sum.

Initialize B , G randomly

- Update B
- $\nabla c(B, G)$
- Update G
- Update C

Repeat until C is small

Article solution - Synthesis

- Comparison of the results in the time domain require phases.
- Original phase – usually good results.
- Generated phase – don't guarantee good results
 - discontinuities at ends of frames.
 - the phase isn't necessarily time aligned.

Article solution - Experiment

- Comparison
ISA, NMF – Euclidean, Divergence, Log
- Choosing 5, 10, 15, 20 components + average
- Automatic clustering with original data as ref.

- SNR:

- m^{th} – ref.
- J^{th} – component.

$$SNR(m, j) = \frac{\sum_{k,t} [Y_m]_{k,t}^2}{\sum_{k,t} \left([Y_m]_{k,t} - [\hat{Y}_j]_{k,t} \right)^2}$$

Article solution - Results

- The difference between the algorithms is statistically significant.
- Temporal continuity term improves the pitches source detection. $\alpha > 0$
- Sparseness doesn't improves the results. $\beta > 0$
- If either α , β are too big the results become poor.

Article solution - Results

TABLE II
SIMULATION RESULTS

algorithm	detection error rate (%)			SNR (dB)		
	all	pitched	drums	all	pitched	drums
ISA	31	29	33	3.6	4.4	1.9
NMF-EUC	28	28	30	6.6	7.9	3.7
NMF-DIV	26	28	23	7.0	8.8	3.5
NMF-LOG	80	90	57	2.3	2.7	2.2
proposed	24	25	22	7.3	9.1	3.6

- NMF-DIV > NMF-EUC > ISA

Article solution - Results

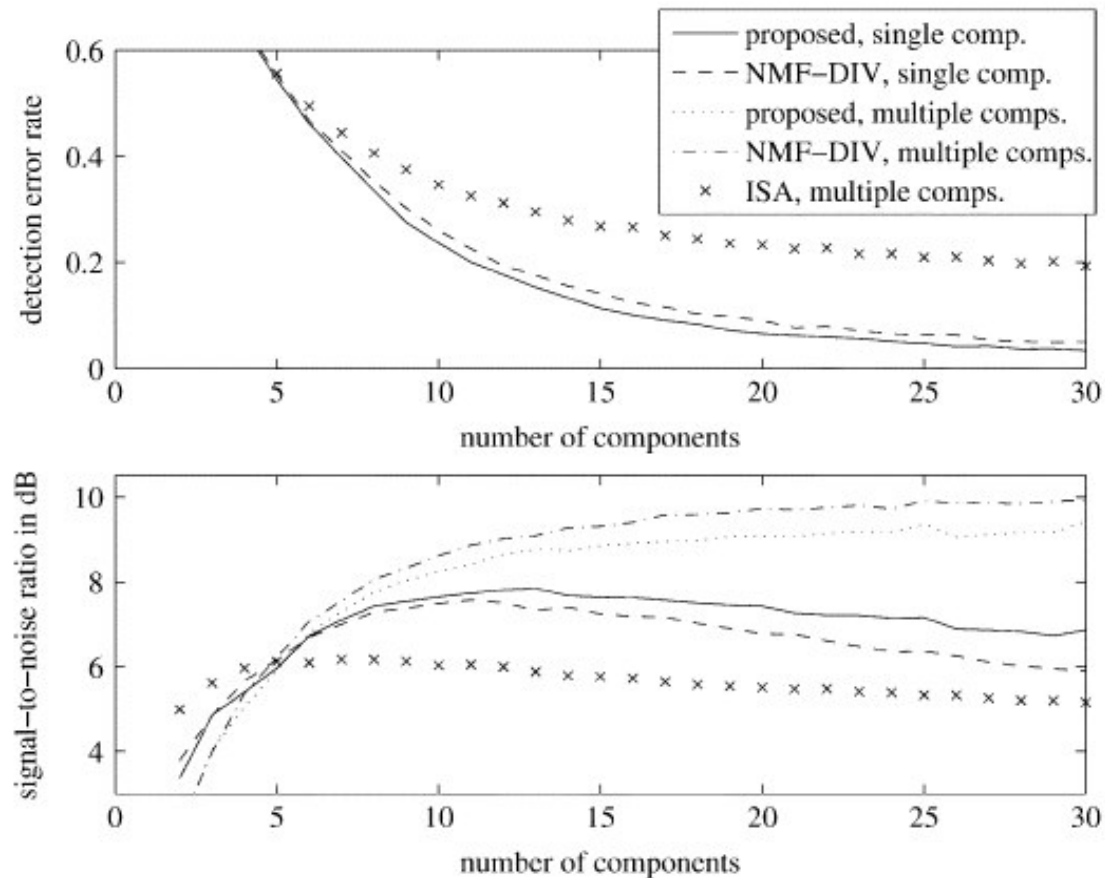


Fig. 3. Illustration of the effect of the component count. “Single comp.” refers to measures where a single component was used to model each source, and “multiple comps.” refers to measures where all the components were clustered using the original signals as references.

Article solution - Conclusions

- Temporal continuity improves the detection of pitched sounds.
- C_t is simple and efficient for the temporal continuity in the cost function.
- The sparseness assumptions wasn't much helpful.
- NMF and the article solution are better than the ISA.
- NMF isn't enough – more assumptions needed.
- The proposed multiplicative rules are efficient for non-negative parameters, but do not guarantee to decrease the cost function.

Article solution - Pro's and Con's

- **Pro's:**

- The article is well written and possible to understand.
- The level of explanation is very good – both of the algorithm and of the results.
- The algorithm itself is possible to understand and possible to duplicate.

- **Con's**

- There is no theoretical explanations to the update rule of G.
- There is a problem with boundaries in the ∇C_t^- term - it is not mentioned how it was solved (set to 0).
- The results are very hard to check as a hole because of the amount of data required.

Article solution - Pro's and Con's

- **Con's:**

- The function $f(x)$ was chosen because it was less sensitive to the sparseness term – that term was of little importance.
- The results are dependant on the comparison method – using the measure:

$$\sum_{k,t} \left(\frac{[Y_m]_{k,t}^2}{[\hat{Y}_j]_{k,t}^2} - 1 + \log \frac{[\hat{Y}_j]_{k,t}^2}{[Y_m]_{k,t}^2} \right)$$

Changes the results – now ISA is much better than the rest of the algorithms.

- Divide by 0: C_r , update B, $\text{grad}(c_r)$

Future Work

1. Different update rules for B , G .
2. Different functions 'f(x)' in the C_S term.

049035 - Digital Speech Processing in Noisy Environments

References

1. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria Tuomas Virtanen. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 3, MARCH 2007
2. M. Goto, "A predominant-f₀ estimation method for real-world musical audio signals: MAP estimation for incorporating prior knowledge about f₀s and tone models," in Proc. Workshop Consistent Reliable Acoust. Cues for Sound Anal., 2001
3. G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 5, no. 5, pp. 1135–1150, Sep. 2004.
4. G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single channel source separation," *J. Mach. Learn. Res.*, vol. 23, pp. 1365–1392, 2003.
5. S. Dubnov, "Extracting sound objects by independent subspace analysis," in Proc. 22nd Int. Audio Eng. Soc. Conf., Espoo, Finland, Jun.2002.
6. J. C. Brown and P. Smaragdis, "Independent component analysis for automatic note extraction from musical trills," *J. Acoust. Soc. Amer.*, vol. 115, pp. 2295–2306, May 2004.
7. P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in IEEE Workshop on Applications of Signal Process. Audio Acoust., New Paltz, NY, 2003, pp. 177–180.
8. S. A. Abdallah and M. D. Plumbley, "An independent component analysis approach to automatic music transcription," in Proc. Audio Eng. Soc. 114th Convention, Amsterdam, The Netherlands, Mar. 2003.
9. T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in Proc. Int. Comput. Music Conf., Singapore, 2003, pp. 231–234.
10. E. Vincent and X. Rodet, "Music transcription with ISA and HMM," in Proc. 5th Int. Symp. Independent Compon. Anal. Blind Signal Separation, Granada, Spain, 2004, pp. 1197–1204.
11. D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Neural Inf. Process. Syst.*, Denver, CO, 2001, pp. 556–562.