E6886 Project Proposal

Yi Luo
UNI: yl3364

When deciding what kind of project should I do, I asked Prof. Wright for advice on two topics: blind source separation with matrix factorization in music signal and recurrent neuron network on singing voice region detection in polyphonic music signal. Finally, I decided to do some research and experiment on the usage of NMF in audio decomposition, especially in separating human singing voice from accompaniment in monaural music recordings.

The problem of singing voice separation is: given a polyphonic music recording (i.e. containing human voice and other instruments in one channel), what we want to get is the pure singing voice. This can be generalized to separate any musical informative source from a polyphonic recording.

NMF-based audio decomposition has been widely studied, and the innovation mainly falls into these aspects:

(1) Consider about different divergence. Although K-L divergence is the one that used in the original NMF paper [1], many other divergences have been studied, such as Itaku-Saito divergence (known as IS-NMF) [2], α-divergence [3], etc.

(2) Adding extra constraints. Since the audio perception ability of human is much more complicatedly than what we could simulate by machine, we could only adopt some simple perception constraints when doing matrix factorization. For example, [4] used some constrains like *loudness index*, [5] used inharmonic constraints in singing voice. Moreover, sparsity is also considered by researchers. In [6] the author claims that "when the spectrum of one source covers partly the spectrum of another, the latter source could be modeled as a sum of the first sound and a residual". Singing singing voice only covers partly of the accompaniment, sparsity could also be used in modeling it.

(3) Combining with other methods. For examples, [7] added Bayesian extensions to NMF, [8] combined long short-term memory RNN with NMF.

What I want to do is to try to add some extra constraints, especially sparsity constraints and perception constraints. Specifically, I want to follow the pitch-based approach (which means that first extract the pitch contour for the human singing voice from the polyphonic music, and use this pitch information as a cue to decompose the matrix), then apply some sparsity constraints (e.g. sparse encoding of harmonics or uncontinuity in spectragram) and perception constraints (e.g. the model of human ear or human perception system) then finally do the NMF. The dataset I want to use is MIR-1K dataset for singing voice separation. I'm still reading papers so maybe I'll change some of the detailed algorithms, but my basic idea is here.

[1] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]//Advances in neural information processing systems. 2001: 556-562.

[2] Févotte C, Bertin N, Durrieu J L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis[J]. Neural computation, 2009, 21(3): 793-830.

[3] Cichocki A, Lee H, Kim Y D, et al. Non-negative matrix factorization with α-divergence[J]. Pattern Recognition Letters, 2008, 29(9): 1433-1440.

[4] Virtanen T O. Monaural sound source separation by perceptually weighted non-negative matrix factorization[J]. Tampere University of Technology, Tech. Rep, 2007.

[5] Duan Z, Zhang Y, Zhang C, et al. Unsupervised single-channel music source separation by average harmonic structure modeling[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2008, 16(4): 766-778.

[6] Virtanen T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2007, 15(3): 1066-1074.

[7] Virtanen T, Godsill S. Bayesian extensions to non-negative matrix factorisation for audio signal modelling[C]//Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008: 1825-1828.

[8] Weninger F, Schuller B, Wöllmer M, et al. Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory[C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5840-5843.

[9]