FACTORS IN FACTORIZATION: DOES BETTER AUDIO SOURCE SEPARATION IMPLY BETTER POLYPHONIC MUSIC TRANSCRIPTION ?

Tiago Fernandes Tavares[*], *George Tzanetakis, Peter Driessen*

University of Victoria
Department of Computer Science
3800 Finnerty Road - Victoria, BC, Canada

ABSTRACT

Spectrogram factorization methods such as Non-Negative Matrix Factorization (NMF) are frequently used as a way to separate individual sound sources from complex sound mixtures. More recently, they have also been used as a first stage for the automatic transcription of polyphonic music. The problem of sound source separation is different (but related) to the problem of automatic music transcription. The output of the first is the separated audio signals corresponding to each sound source, whereas the output of the second is a symbolic representation/music score that encodes the discrete pitches/notes that are played and when they are played. Many variations of factorization methods have been proposed. Two important design choices are the way spectra are represented and what distance measures are used to compare them in the optimization used for factorization. A common assumption has been that a variant that yields better signal separation will result in better automatic transcription. In this work, we investigate experimentally this question and show that this relationship is not necessarily true.

Index Terms— Non-negative matrix factorization, Music Transcription, Beta-Divergence, Sound Source Separation.

1. INTRODUCTION

Automatic Music Transcription (AMT) is the task of detecting musical notes in an audio signal. Musical notes are described by their *onset* (time instant when the note is triggered), *offset* (time instant when the note is damped) and *pitch* (describing *which* note is being played). Other information, such as velocity, is often neglected. Some of the current applications for AMT systems are in query-by-content databases [1], tutoring software [2] and musical analysis [3].

Earlier work in AMT was mostly based on multiple pitch estimation methods. More recently several AMT systems based on spectrogram factorization methods that originated in audio source separation have been proposed. These systems rely on the assumption that the short time spectrum of an audio signal is a linear combination of the spectra corresponding to the individual notes that are active during the time period over which the spectrum is computed:

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\phi}, \tag{1}$$

where $\boldsymbol{x}$ is the short-time power spectrum of the mixture audio signal at a particular time, $\boldsymbol{B}$ is a matrix where each column is a prototype spectrum corresponding to a particular note, $\boldsymbol{a}$ is an activation vector that has high values for the notes that are active and low values for notes that are not active during that time, and $\boldsymbol{\phi}$ is the approximation error. When all the short-time spectra of the audio signal at different times are considered together, they form the spectrogram $\boldsymbol{X}$. The result over time is the activation matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_Q]$. This activation matrix can then be thresholded, yielding the desired AMT output (similar to a “piano roll” format). The activation matrix can be obtained by using Non-Negative Matrix Factorization (NMF), which aims to minimize a divergence $d(\boldsymbol{X}|\boldsymbol{B}\boldsymbol{A})$ between $\boldsymbol{X}$ and $\boldsymbol{B}\boldsymbol{A}$ with the constrain of non-negativity for all matrix elements.

NMF for audio signals was originally applied for audio source separation where the desired output is the audio signals corresponding to the individual sound sources in a mixture. In AMT, each individual note is considered as a different audio source, so its activity level (i.e., how loud it is sounding) at each moment is obtained using NMF. Due to this proximity, a common assumption underlying factorization approaches for AMT is that algorithms that obtain good sound source separation will also result in better AMT. In this work, we experimentally investigate this relationship using different variants of the common NMF algorithm. We consider the optimistic scenario where $\boldsymbol{B}$ is known and is not estimated during the factorization (this enables real time calculation of $\boldsymbol{A}$). The results show that the relationship between source separation and transcription performance is weak.

2. RELATED WORK

NMF has been used in audio signal processing for solving under-determined source separation problems [4, 5, 6] as well

[*] Thanks CAPES-BRAZIL for funding.

as AMT [7]. In Smaragdis and Brown's work [7], both $\boldsymbol{B}$ and $\boldsymbol{A}$ were obtained by minimizing $\|\boldsymbol{X} - \boldsymbol{BA}\|$. A common assumption when using NMF is that a technique that yields the best $\boldsymbol{A}$ will also yield the best approximation $\boldsymbol{Y} = \boldsymbol{BA}$. Hence, different variants of the basic NMF algorithm have been used to improve how well $\boldsymbol{X}$ can be approximated using the base $\boldsymbol{B}$. One of these techniques is changing the spectral representation used in the factorization process.

By using the power spectrum ($|\boldsymbol{y}|^2$), low-energy components become less relevant in the approximation, but the magnitudes of the high-energy components tend to exhibit larger variations [8, 9]. On the other hand, a logarithmic representation ($\log(1 + |\boldsymbol{y}|)$) can be more stable, while giving more importance to low-energy components [10]. The magnitude of the DFT ($|\boldsymbol{y}|$) has also been used in AMT [11, 12], and can be viewed as a tradeoff between the advantages and disadvantages of the logarithmic and power spectrum. It is possible to exploit the fact that the columns of $\boldsymbol{B}$ are roughly a frequency-domain shift of each other [13]. Also, it is possible to assign more than one base vector for each note [14] improving the approximation. Changing the divergence measure that is minimized can yield better transcription results. Dessein *et. al* [9] used the beta divergence [15], defined as:

$$\epsilon_\beta(x|y) = \begin{cases} \frac{(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1})}{\beta(\beta-1)}, & \beta \in \mathbb{R}\backslash\{0,1\} \\ x - \log\frac{x}{y} + (x-y), & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1, & \beta = 0. \end{cases} \tag{2}$$

All values of β lead to the same global minimum ($\epsilon_\beta(x|x) = 0$), but the behaviour of the divergence for $x \neq y$ is different. These different variants can lead to different inexact approximations. We experimentally investigate how these variants affect performance using both source separation and AMT evaluation metrics.

3. METHODOLOGY

Our experiments aimed at comparing the performance difference caused by changing the factorization method in an automatic transcription algorithm and a source separation process. For such, it is necessary to have a dataset consisted of audio files and time-aligned ground-truth. The dataset we used was created utilizing the MIDI files used by Poliner and Ellis [16]. The files were downloaded from a MIDI Database [17] and rendered using the same piano note samples that were used to obtain the base matrix. The samples were downloaded from the Iowa Musical Instrument Samples database [18]. This is a best case scenario where there is prior knowledge of the sound source material and the challenge is to recover the activations of the polyphonic mixture. In total, the database consists of 24 files containing 174 minutes of polyphonic piano Classical music, which has showed to suffice for obtaining statistically valid results.

The base matrix $\boldsymbol{B}$ was obtained by averaging the spectrograms of the first 250 milliseconds from the recordings of each individual piano note. This corresponds to 10 frames considering frames of 2048 samples with a hop size of 1024 and a sampling rate of 44100 Hz. The higher harmonics of piano notes decay more quickly. Therefore, this time-domain pruning avoids using parts of the note with low harmonic content.

We avoided using approaches that allow the adaptation of the base matrix $\boldsymbol{B}$. In this kind of approach, the relationship between each base vector and each note could be lost [7]. Therefore, the problem of assigning bases to notes would arise, and would represent a different parameter to be analyzed.

Four values for β were considered. When $\beta = 2$ (Euclidean distance), the result is a minimum squared error approximation. When $\beta = 1$ (Kullback-Liebler divergence), a minimum mutual entropy solution is obtained. The Itakura-Saito divergence ($\beta = 0$) corresponds to the Gaussian, maximum-likelihood approximation. Finally, the value $\beta = 0.5$ was used as it has been shown to give good results in the context of AMT [9]. Three different representations for the spectrogram were used, as discussed in Section 1: the logarithmic ($\log(1 + |\boldsymbol{x}|)$), the magnitude ($|\boldsymbol{x}|$) and the power ($|\boldsymbol{x}|^2$). The activation vector $\boldsymbol{a}$ was obtained for each audio frame by means of an iterative algorithm [19].

A symbolic transcription was obtained by thresholding the $\boldsymbol{A}$ matrix and excluding from the final results all notes whose duration was less than 50 ms. The transcription results were analyzed using the evaluation metrics used in the Music Information Retrieval Exchange (MIREX) [20], that is, the Recall (R; true positives divided by the total number of notes in ground truth), Precision (P; true positives divided by the total number of yielded notes) and, finally, their harmonic mean called the F-Measure ($F = 2RP/(R + P)$). A note is considered correct if, when compared to a ground-truth annotation, its pitch is within half a semitone, its onset is within 50 ms, and its duration does not deviate by more than 20% of the reference. Note offsets are often considered not as important as onsets [20] and can be excluded from the evaluation. For each case, the threshold applied to $\boldsymbol{A}$ was the one that maximized the F-Measure over the whole database, that is, the same threshold was used for all pieces.

To measure the accuracy of the separation process, we evaluated how well the model yielded by the factorization model could reconstruct the original signal. Using the activation and the base matrix, two different reconstructions were evaluated: the spectrogram reconstruction and the audio signal reconstruction. The spectrogram reconstruction, for each frame, consisted of calculating the approximation $\boldsymbol{y} = \boldsymbol{Ba}$ for each frame, yielding a spectrogram $\boldsymbol{Y}$. The original audio file was reconstructed frame-to-frame by performing overlap-and-add using the inverse DFT of vectors whose magnitude values were taken from $\boldsymbol{Y}$ and the phase values from $\boldsymbol{X}$, a standard phase-vocoder technique [21].

To evaluate the spectrogram reconstruction, the measures proposed by Vincent *et. al* [22] were used: the Signal-to-Interference Ratio (SIR), the Signal-to-Artifacts Ratio (SAR) and the Signal-to-Distortion Ratio (SDR). In this evaluation, each frequency bin in the spectrogram is considered as a different source. Therefore, the SIR represents how much each frequency interferes with each other, the SAR highlights errors caused by artifacts and the SDR gives a general idea of the reconstruction error considering both noise, interference and artifacts. To evaluate the time-domain signal reconstruction, only the SDR was used, as it is a monaural signal so the other measures do not apply.

We calculated the Spearman (rank) correlations between each one of the reconstruction measures and the respective F-Measures. In addition to the correlation values, a P-Value was also calculated, representing the probability that the two compared measurements are not correlated. This value is obtained using a T-Student test, and, when it is lower than 5%, the compared measures may be considered correlated. The correlations were obtained considering, first, the whole set of data. Then, the value of β was fixed and the results using different pieces and spectral representations were considered. This aims to show the influence of the spectral representation in the final results. Two other similar partitions were considered, fixing the spectral representation and the piece.

4. EXPERIMENTAL RESULTS

The average performance measures (F-Measures and distortion measures) for each spectral representation and β are reported in Tables 1, 2 and 3. From these results, it is possible to detect some trends. In Table 1, for all values of β the magnitude and the power spectrum representations yielded, respectively, the best and the worst results. It can also be seen that the results when considering only the onsets and when also considering offsets are greatly correlated. Table 2 shows that the only distortion measure that was clearly improved by using the magnitude representation was the SIR – the SDR and the SAR were, in general, best when using the logarithmic representation. As shown in Table 3, the worst signal reconstruction distortion values were obtained when using the magnitude representation. The factorization using $|\boldsymbol{X}|^2$ and $\beta = 0.0$ did not yield an activation matrix with a meaningful outcome for transcription, but the results in tables 2 and 3 show that this factorization can be used to reconstruct the original signal with less distortion.

The rank correlation over the entire the dataset was calculated, yielding the results shown in Table 4. The effects of the value of β were evaluated by calculating the correlation for each value of β separately, yielding the results shown in Table 5. A similar experiment was performed considering each different spectral representation, yielding the results shown in Table 6. Last, the results for each individual piece were considered, so that the effect of the musical complexity could be evaluated. For this test, no significant or strong correlations were found, so the result table was omitted.

Table 1. Average F-Measures (considering onset only/onset and offset) for each spectral representation and β.

β	$\log(1+\lvert\boldsymbol{X}\rvert)$	$\lvert\boldsymbol{X}\rvert$	$\lvert\boldsymbol{X}\rvert^2$
2.0	0.74/0.42	0.79/0.47	0.68/0.28
1.0	0.80/0.51	0.84/0.54	0.72/0.31
0.5	0.78/0.50	**0.86/0.60**	0.76/0.32
0.0	0.72/0.45	0.80/0.58	0.00/0.00

Table 2. Average spectral reconstruction distortion (SDR/SIR/SAR), in dB, for each representation and β.

β	$\log(1+\lvert\boldsymbol{X}\rvert)$	$\lvert\boldsymbol{X}\rvert$	$\lvert\boldsymbol{X}\rvert^2$
2.0	2.6/2.83/18.4	5.8/6.4/18.8	-12.0/-11.7/13.0
1.0	4.6/5.0/18.3	8.2/8.9/ **19.6**	-14.8/-13.7/7.7
0.5	6.0/6.5/18.0	**8.9**/ **9.9** / 18.0	-21.4/-18.8/3.7
0.0	6.7/7.4/16.6	2.2/5.7/7.1	-18.9/-16.3/5.2

Table 3. Average time-domain reconstruction distortion (SDR-a), in dB, for each representation and β.

β	$\log(1+\lvert\boldsymbol{X}\rvert)$	$\lvert\boldsymbol{X}\rvert$	$\lvert\boldsymbol{X}\rvert^2$
2.0	**-5.3**	-12.7	-11.67
1.0	-5.9	-13.0	-11.7
0.5	-6.0	-12.9	-9.7
0.0	-6.7	-15.2	-8.0

Table 4 shows that, for the whole dataset, all correlations between distortion measures and the transcription results are significant, but weak. It can be seen that the SAR is more important when finding only onsets than when finding onsets and offsets. Interestingly, the results show a negative correlation between the SDR-a and both F-Measures, which means a better signal reconstruction tends to indicate worse transcriptions. These results are confirmed in Table 5, except for the cases where $\beta = 0.0$. In these cases, all calculated correlations are negative and, interestingly, very weak in for the SDR and SIR when considering the transcription evaluation with onsets and offsets. Very weak correlations were also found for the SDR-a when $\beta = \{2.0, 1.0\}$. The greatest correlation value were, in general, found for the SIR and the SDR, but they were consistently lower than 0.65, which indicates a weak to average correlation. Table 6 shows a similar set of results. Weak correlations are consistently found for SDR, SIR and SAR, and negative correlations are found for SDR-a.

In general, the distortion measures show only a weak to average correlation with the F-Measure. It can be noted that the SIR and SDR are more correlated to the transcription results than the SAR. However, the difference between them is lower when considering only the detection of onsets. Negative correlations were obtained for the SDR-a, regardless of

Table 4. Spearman (rank) correlation between the F-Measure and different distortion measures considering whole dataset.

Evaluation	SDR	SIR	SAR	SDR-a
Onset only	**0.37** (P=3 $\times 10^{-10}$)	**0.37** (P=2 $\times 10^{-10}$)	0.31 (P=2 $\times 10^{-7}$)	-0.48 (P=5 $\times 10^{-7}$)
Onset and offset	0.43 (P=9 $\times 10^{-14}$)	**0.48** (P=$\times 10^{-16}$)	0.18 (P=0.0019)	-0.27 (P=0.006)

Table 5. Spearman (rank) correlation between the F-Measure and different distortion measures considering fixed values of β.

Evaluation	β	SDR	SIR	SAR	SDR-a
Onset only	2.0	0.51 (P=3 $\times 10^{-6}$)	0.51 (P=4 $\times 10^{-6}$)	0.45 (P=5 $\times 10^{-5}$)	-0.29 (P=0.012)
	1.0	**0.54** (P=9 $\times 10^{-7}$)	0.53 (P=1 $\times 10^{-6}$)	**0.54** (P=8 $\times 10^{-5}$)	-0.23 (P=0.044)
	0.5	0.44 (P=9 $\times 10^{-5}$)	0.44 (P=8 $\times 10^{-5}$)	0.33 (P=0.004)	-0.47 (P=2 $\times 10^{-5}$)
	0.0	-0.44 (P=0.002)	-0.50 (P=0.0002)	-0.31 (P=0.02)	-0.65 (P=5 $\times 10^{-7}$)
	β	SDR	SIR	SAR	SDR-a
Onset and Offset	2.0	0.51 (P=3 $\times 10^{-6}$)	0.52 (P=2 $\times 10^{-6}$)	0.25 (P=0.02)	-0.05 (P=0.64)
	1.0	0.58 (P=8 $\times 10^{-8}$)	0.58 (P=5 $\times 10^{-8}$)	0.37 (P=0.001)	-0.08 (P=0.48)
	0.5	0.62 (P=3 $\times 10^{-9}$)	**0.64** (P=1 $\times 10^{-9}$)	0.3 (P=0.008)	-0.29 (P=0.012)
	0.0	-0.18 (P=0.17)	-0.11 (P=0.42)	-0.40 (P=0.004)	-0.49 (P=0.0003)

Table 6. Spearman (rank) correlation between the F-Measure and different distortion measures considering fixed representations.

Evaluation	β	SDR	SIR	SAR	SDR-a
Onset only	$\log(1+\|\boldsymbol{X}\|)$	-0.041 (P=0.68)	-0.06 (P=0.54)	0.31 (P=0.001)	-0.48 (P=5 $\times 10^{-7}$)
	$\|\boldsymbol{X}\|$	**0.42** (P=1 $\times 10^{-5}$)	0.39 (P=7 $\times 10^{-5}$)	0.26 (P=0.10)	-0.36 (P=0.0002)
	$\|\boldsymbol{X}\|^2$	-0.31 (P=0.006)	-0.27 (P=0.019)	-0.38 (P=0.0008)	-0.15 (P=0.2)
	β	SDR	SIR	SAR	SDR-a
Onset and Offset	$\log(1+\|\boldsymbol{X}\|)$	**0.31** (P=0.001)	0.30 (P=0.002)	0.03 (P=0.74)	-0.27 (P=0.006)
	$\|\boldsymbol{X}\|$	0.23 (P=0.019)	**0.31** (P=0.001)	-0.12 (P=0.21)	-0.26 (P=0.0095)
	$\|\boldsymbol{X}\|^2$	0.06 (P=0.61)	0.11 (P=0.32)	-0.18 (P=0.11)	-0.11 (P=0.35)

the positive correlations for SDR. The SDR is obtained by averaging the SDR amongst all frequency domain coefficients. For the SDR-a, the high energy components have a greater impact on the final results. Thus, high energy components are estimated with greater noise than low energy components. Therefore, it is not important to estimate the exact amplitude of the high-energy components of the audio signal, as long as the estimations are good enough to result in the detection of the corresponding note events.

Although the obtained correlation values are low to average, the fact that the correlation between these measures is positive and significant means that they have some relation to transcription performance. Therefore, the distortion measures, especially the SIR, may provide clues regarding factorization methods that can yield good transcription results. However, these measures alone can not provide a definitive conclusion on which variant/method is better. The results also show that the factorization requirements for detecting only onsets are slightly different than the requirements for detecting both onsets and offsets. When only onsets are considered, the presence of artifacts has a greater impact on the final results, but if offsets are also considered, the distortion in each frequency band and the interference between the different frequency bands becomes more relevant. The next section presents some conclusions and guidelines for the use of factorization methods for audio source separation and automatic music transcription.

5. CONCLUSIONS

This paper discusses the correlation between the transcription accuracy and the separation abilities of beta-divergence non-negative factorization methods. Several automatic music transcribers based on variants of NMF were built. They differ in the spectral representation (logarithmic, magnitude and power) and the values of β (0, 0.5, 1, 2). Following the factorization process, a threshold was applied, yielding discrete notes. The transcriptions yielded by each algorithm were evaluated using the F-Measure (as defined in the MIREX [20]), and the separation abilities were measured using the SDR, SIR and SAR of the spectrogram reconstruction and the SDR of the audio signal reconstruction, as defined by Vincent *et. al* [22]. The Spearman (rank) correlation between the F-Measures and each of the distortion measures was calculated. To the best of our knowledge this is the first detailed experimental investigation of factorization methods that considers both audio source separation and automatic music trancsription.

The separation capabilities of the algorithm appear to have only a partial effect on the transcription quality, with a correlation lower than 0.5. The greatest correlation value was found for the SIR measure. Thus, the most important characteristic of the factorization is to maintain the orthogonality between the analyzed frequency components. If the note offsets are not considered in the evaluation, then the importance of the SAR ratio increases. Interestingly, the correlation between the F-Measures and the SDR ratio in the time-domain audio reconstruction is negative, that is, a closer approximation of the audio signal as a linear combination of base vectors does not imply a better transcription.

We conclude that the obtained F-Measures are related to characteristics that are not necessarily the ones that will yield a better audio source separation. Therefore, the separation abilities of an algorithm should not be used as a conclusive argument on why it should be used in music transcription. In the absense of other information a good choice for both tasks is to use a magnitude spectrum representation and a β value of 0.5. Future work should aim at finding what are the characteristic of a factorization that are changed by using particular signal representations for β values, and how it correlates with the F-Measure. Such characteristics are important because if they could be calculated directly from the factorization results then they could help design AMT systems that optimize a more problem specific objective function.

6. REFERENCES

[1] J. Li, J. Han, Z. Shi, and J. Li, "An efficient approach to humming transcription for query-by-humming system," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 8, Oct. 2010, pp. 3746 –3749.

[2] J. Yin, Y. Wang, and D. Hsu, "Digital violin tutor: an integrated system for beginning violin learners," in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 976–985. [Online]. Available: http://doi.acm.org/10.1145/1101149.1101353

[3] C. C. Liem, A. Hanjalic, and C. S. Sapp, "Expressivity in musical timing in relation to musical structure and interpretation: a cross-performance, audio-based approach," in *AES 42nd International Conference*, Jul. 2011.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1066 – 1074, Mar. 2007.

[5] M. N. Schmidt, "Single-channel source separation using non-negative matrix factorization," Ph.D. dissertation, Technical University of Denmark, Nov. 2008.

[6] A. Lefevre, F. Bach, and C. Fevotte, "Semi-supervised nmf with time-frequency annotations for single-channel source separation," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, Oct. 2012.

[7] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, Oct. 2003, pp. 177 – 180.

[8] S. A. Abdallah and M. D. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Conf. Music Inf. Retrieval (ISMIR'04)*, Barcelona, Spain, 2004.

[9] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, Aug. 2010, pp. 489–494.

[10] B. Niedermayer, "Non-negative matrix division for the automatic transcription of polyphonic music," in *Proceedings of the ISMIR*, 2008.

[11] G. Grindlay and D. Ellis, "Multi-voice polyphonic music transcription using eigeninstruments," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, oct. 2009, pp. 53 –56.

[12] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 538 –549, Mar. 2010.

[13] H. Kirchhoff, S. Dixon, and A. Klapuri, "Shift-variant non-negative matrix deconvolution for music transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, Kyoto, Japan, Mar. 2012, pp. 125–128.

[14] ——, "Multi-template shift-variant non-negative matrix deconvolution for semi-automatic music transcription," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, Oct. 2012.

[15] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," *Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep*, 2001.

[16] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong, “Melody transcription from music audio: Approaches and evaluation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1247 –1256, May 2007.

[17] mididatabase.com, “The midi database,” http://mididatabase.com/.

[18] University of Iowa, “Musical Instrument Samples,” ”http://theremin.music.uiowa.edu/MIS.html”, 2005.

[19] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the beta-divergence,” *CoRR*, vol. abs/1010.1763, 2010.

[20] S. J. Downie, “The music information retrieval evaluation eXchange (MIREX),” *D-Lib Magazine*, vol. 12, no. 12, Dec. 2006.

[21] J. L. Flanagan and R. M. Golden, “Phase vocoder,” *Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[22] E. Vincent, C. Fevotte, and R. Gribonval, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.