2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SATANDER, SPAIN

OPTIMAL COST FUNCTION AND MAGNITUDE POWER FOR NMF-BASED SPEECH SEPARATION AND MUSIC INTERPOLATION

Brian King[1,4] *Cédric Févotte*[2] *Paris Smaragdis*[3,4]

[1] Department of Electrical Engineering, University of Washington
[2] CNRS LTCI, Télécom ParisTech
[3]Department of Computer Science, University of Illinois at Urbana-Champaign
[4]Advanced Technology Labs, Adobe

ABSTRACT

There has been a significant amount of research in new algorithms and applications for nonnegative matrix factorization, but relatively little has been published on practical considerations for real-world applications, such as choosing optimal parameters for a particular application. In this paper, we will look at two applications, single-channel source separation of speech and interpolating missing music data. We will present the optimal parameters found for the experiments as well as discuss how parameters affect performance.

Index Terms— Nonnegative matrix factorization, source separation, spectrogram interpolation

1. INTRODUCTION

Nonnegative matrix factorization (NMF) has become a popular area of research and has been applied in many diverse fields, including audio. Within audio, it has been applied in a variety of tasks, including single-channel source separation [1], interpolation of missing audio data [2], bandwidth expansion [3], polyphonic transcription [4], and multi-source and noise-robust speech recognition and dynamic time warping [5, 6]. There has been a significant amount of research on these topics, with many new algorithms and parameters being proposed. While it is exciting to see so much work in this field, it has become challenging to choose an NMF algorithm for a particular application because there are many papers proposing different cost functions and parameters that purport to be the best. At the moment, it can take a significant amount of time to search through the literature and run experiments to find the best parameters for the chosen application. The goal of this paper is to help address this challenge. In this paper, we focus on two popular applications, single-channel separation of speech and interpolation of missing music data. For each, we ran experiments with many different NMF models and parameters. In our data analysis, we will discuss how parameters affect performance, provide explanations and hypotheses about the performance trends we observe, as well as present our findings for parameters that perform optimally overall, in a fashion similar to the parameter analysis for NMF-based musical source separation in [7] and NMF-based multi-pitch estimation in [8]. Our goals are to show readers what models and parameter values work well, as well as help develop an intuition for how parameters affect performance. And although we focus on two applications in this paper, we hope that this knowledge of how parameters affect performance will lead to an ability to understand, and even predict, how parameter choice will affect the performance of different applications.

This work is supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and Applications of Nonnegative Matrix Factorization).

2. BACKGROUND

The NMF algorithm decomposes an observed matrix into a product of a basis and weight matrix,

$$\begin{aligned} &V \approx \hat{V} = WH, \\ &\text{where } V \in \mathbb{R}^{F\times T}, W \in \mathbb{R}^{F\times K}, H \in \mathbb{R}^{K\times T}, \\ &\text{and } V_{ft}, W_{fk}, H_{kt} \geq 0 \text{ for all elements} \end{aligned} \tag{1}$$

The key differentiator of NMF from other matrix factorization models, such as independent component analysis or principal component analysis is that all of the elements of the observed data (V), basis (W), and weight matrix (H) are nonnegative and real. This means that the basis vectors combine additively, never canceling each other out. When applied to audio, the observed matrix is most often the magnitude or power spectrogram, but can also be other nonnegative time-frequency representations [9]. When applied to a time-frequency representation, the columns of W correspond to spectral basis vectors and the columns of H indicate the weighting of those basis vectors within a particular time window. In audio applications, the nonnegativity constraint on the matrices can often result in meaningful basis vectors, which can be helpful for analysis or processing. For example, NMF analysis of a piano excerpt can find basis vectors that each correspond to the spectra of individual notes [4], and NMF analysis of speech can result in phoneme-like bases [1].

Although the basic NMF algorithm is fairly simple, there are many variations that have been proposed in the research literature, including differences in cost function, magnitude exponent, number of bases, and window length. In this paper, we will explore how these affect performance, as well as disclose the parameters we found to work best overall. We will now discuss these parameters and provide examples and intuition of how they affect performance. NMF uses an iterative algorithm to updates the basis and / or weight matrices to minimize the given cost function. Thus, different cost functions can produce significantly different results. The most popular cost functions are the squared Frobenius norm [10],

$$d(V|\hat{V})_{FRO} = \sum_{f,t} \parallel V_{ft} - \hat{V}_{ft} \parallel^2 \tag{2}$$

978-1-4673-1026-0/12/$31.00

the Kullback-Leibler (KL) divergence [10],

$$d(V|\hat{V})_{KL} = \sum_{f,t} V_{ft} \log \frac{V_{ft}}{\hat{V}_{ft}} - V_{ft} + \hat{V}_{ft} \quad (3)$$

and the Itakura-Saito (IS) divergence [11]

$$d(V|\hat{V})_{IS} = \sum_{f,t} \left(\frac{V_{ft}}{\hat{V}_{ft}} + \log \frac{V_{ft}}{\hat{V}_{ft}} - 1 \right) \quad (4)$$

The squared Frobenius norm, as it minimizes the squared error of the estimate, is sometimes criticized for audio applications because it places too high an importance on modeling higher-energy components, often at the expense of lower-energy components, which are often still important for audio quality and auditory perception [11]. In contrast, the KL divergence places a more equal emphasis on components with higher and lower energy. And finally, the IS divergence has a completely scale-invariant cost function, meaning

$$d(V|\hat{V})_{IS} = d(\alpha V|\alpha\hat{V})_{IS}, \text{ for } \alpha > 0 \quad (5)$$

These cost functions are all special cases of the beta-divergence, where β's of $2, 1$, and 0 correspond to the squared Frobenius norm, KL divergence, and IS divergence, respectively [12]. β values between and beyond these three values can be used as well, and will be explored and discussed in the paper.

The second parameter we will be examining is the magnitude exponent of the observation, which raises every element of the STFT magnitude $|X|$ to the power p,

$$V_{ft} = |X_{ft}|^p, \text{for } f = 1, ..., F, t = 1, ..., T \quad (6)$$

The most common values of p are 1 and 2, which correspond to the magnitude and power spectra of the signal. However, any value of $p > 0$ can be used for any of the cost functions.

Deciding on the number of basis vectors to use is essential for good performance. Choosing too small a number can result in the basis vectors not being able to approximate the data well, but too many can result in over-fitting the data. The final parameter we will be discussing is the window length. If a window is too short, then there may not be enough spectral differentiation between the sources, causing poor source separation and interpolation. But if the window is too long, then the spectra of the signal will be less stationary within a window, which will also result in poor performance.

3. SPEECH SEPARATION

3.1. Theory

The parts-based decomposition in the NMF algorithm lends itself well to source separation. If it is known that disjoint sets of spectral basis vectors correspond to different sources, then the enhanced signal for source s can be synthesized by multiplying together the basis vectors and weights corresponding to that source,

$$\hat{V}_{sft} = \left(\sum_{k \in K_s} W_{fk} H_{kt} \right) \quad (7)$$

where K_s is the set of spectral basis vectors corresponding to source s. This estimate is used to calculate the time-frequency weights for filtering the original, complex-valued STFT of the mixed signal,

$$Y_{sft} = \left(\frac{\sum_{k \in K_s} W_{fk} H_{kt}}{\sum_{k=1,...,K} W_{fk} H_{kt}} \right) X_{ft} = \left(\frac{\hat{V}_{sft}}{\hat{V}_{ft}} \right) X_{ft} \quad (8)$$

This method ensures that the decomposition is lossless.

One challenge of source separation is how to acquire disjoint sets of basis vectors that represent the different sources well. We will be using the most straightforward method for our speech separation experiments, the fully-supervised copy-to-train method [13, 14]. In this method, we will use the magnitude of the STFT of the training data (raised to the magnitude exponent p), with the training data being other utterances from that same speaker found in the mixture. The idea is that the speech from the spectra of the training data will be similar enough to that of the speech in the test data that it can be estimated well by the training data. The copy-to-train method is repeated for all the speakers in the mixture. Other methods include the fully-supervised factor-to-train, semi-supervised, and unsupervised training methods [14]. We chose the fully-supervised method because we had training data for all the speakers. One known method to increase source separation performance is to eliminate all frames with energy below a threshold, because nearly silent frames do not represent the important speech well [13]. Because of this, we set the threshold at 40 dB below the highest-energy frame of each speaker's training data. We chose copy-to-train over factor-to-train for a few reasons. First of all, if the training data is representative of the test data, the former method is guaranteed to result in meaningful basis vectors, while the other methods will not necessarily do so if a poor number of basis vectors was chosen for factorizing the training data. Both too many and too few basis vectors can result in less meaningful basis vectors and poor separation. Also, it was helpful to eliminate any unnecessary variables to focus on analyzing the other variables to make the results clearer. So in these experiments, we will see how cost function, magnitude exponent, and window length affect performance.

We will use the blind source separation evaluation (BSS Eval) toolkit for analyzing and comparing source separation results [15]. This method measures the source-to-interference ratio (SIR), source-to-artifact ratio (SAR) and source-to-distortion ratio (SDR). The SIR measures how much of the interfering sources are left in the separated signal. The SAR measures how much energy is in the signal that is not part of either the target or interfering signals. The SDR combines the SIR and SAR into one measurement. The SIR results are computed by finding the difference, in dB, of the SIR of the enhanced signal from the SIR of the original mixed signal. The SDR is computed in the same fashion. The SAR is computed simply by the SAR of the enhanced signal. Since there are no artifacts in the original mixture, its SAR is $+\infty$ dB. Thus, if we used this value in the calculation, the resulting SAR difference would always be $-\infty$ dB, which is not helpful for comparisons. In other words, the baseline SIR and SDR are calculated from the original mixture. We will refer to these measurements as relative SIR, absolute SAR, and relative SDR.

3.2. Experiments and Results

We had the following goals with our single-channel speech separation experiments. First of all, we wanted to see how parameters (cost function, magnitude exponent, and window length) affected performance. We also wanted to test with a variety of target-to-masker ratios (TMR, where the target is the wanted signal and the masker is the unwanted signal) and speaker combinations to determine whether or not there was an optimal set of parameters for all scenarios, and hypothesize why. The test signals consist of two utterances from different speakers from the TIMIT database [16], sampled at 16 kHz. In order to maximize signal overlap in time, we truncated the length of the mixture to the shorter of the two utterances. We then mixed

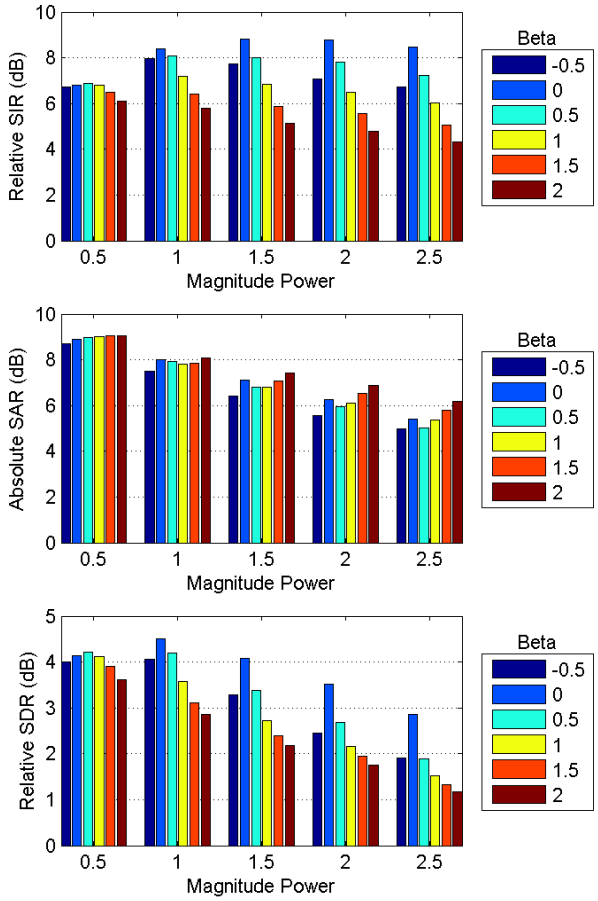

Fig. 1. BSS Eval results illustrating how β and magnitude exponent affect performance (averaged over all experiments with 0 dB TMR and 512-point (32 ms) window). Note: bigger bars are better.

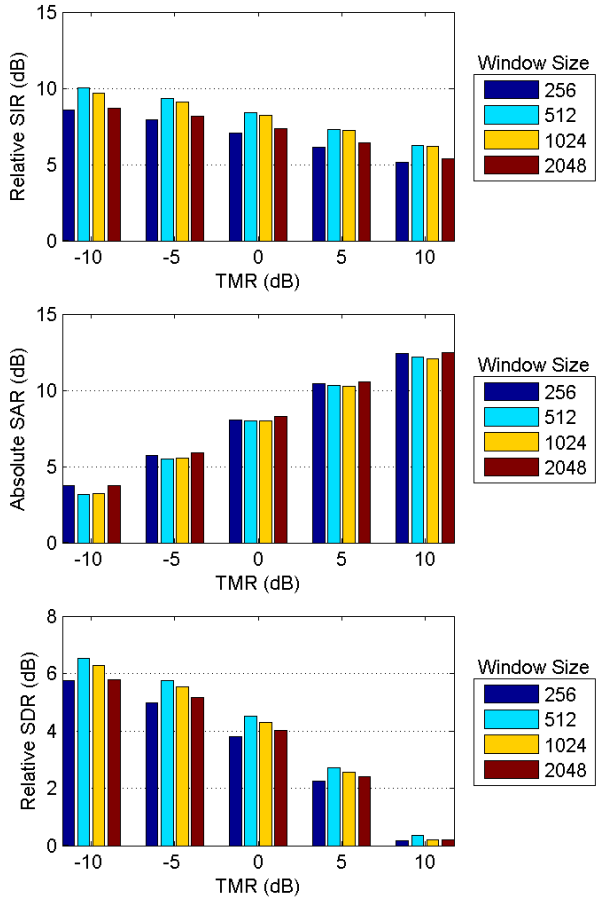

Fig. 2. BSS Eval results illustrating how window size and TMR affect performance (averaged over all experiments with β of 0 and magnitude exponent of 1). With a 16 kHz sampling rate, windows of 512, 1024, 2048, and 4096 points correspond to lengths of 32, 64, 128, and 256 ms, respectively.

the utterances at -10, -5, 0, +5, and +10 dB TMR. There were eight speaker combinations of mixed gender, four of female only, and four of male only. For each experiment, we measured the BSS Eval measurements for both speakers in each experiment. Because of this, we had eight target/masker combinations each of F/F, M/M, F/M, and M/F. Each speaker in the TIMIT database has two sentences common to the other speakers and eight unique sentences. We used only the unique sentences. For a given speaker combination, we used seven of the eight sentences for training and the other for the test mixture. For each combination of speakers, we used two different combinations of test and training signals. So for each set of parameters at each TMR, we ran 24 experiments, resulting in 48 BSS Eval measurements.

In our analysis of the results, we found that one set of parameters performed best overall for our speech separation task. The optimal parameters were a β of 0, a magnitude exponent of 1, and a window size of 512 points (32 ms). We have included some figures that illustrate how parameters affect performance.

3.2.1. Magnitude power and β

In Figure 1, we have plotted the BSS Eval results for varying magnitude powers and the β's of the cost function, with a constant window size of 512 (32 ms) and TMR of 0 dB, averaged over all test sentences. We see that the relative SIR is highest when β equals 0 (Itakura-Saito divergence) and magnitude exponent equals 1.5 and 2. A β of 0 and a magnitude exponent of 2 correspond to the generative model of superimposed Gaussian components advocated in [11]. In the SAR results, we see that as the magnitude exponent increases, the absolute SAR decreases. This is logical because if the magnitude exponent were 0, the separated signals would simply be identical signals that would be half the amplitude of the original signal. There would be no artifacts in that signal, but the SIR improvement would also be 0 dB. As the magnitude exponent increases, more filtering takes place, which can lead to more artifacts and thus a lower SAR. Within a magnitude power, β values of 0 (Itakura-Saito divergence) and 2 (squared Frobenius norm) have the highest absolute SAR, though their SAR values vary less within a given magnitude power than between different powers. The relative SDR, which incorporates both relative SIR and SAR, is maximized when β is 0 and the magnitude exponent is 1.

3.2.2. Window size and TMR

In Figure 2, we see how TMR and window size affect performance. The results plotted are with the optimal parameters discussed in the

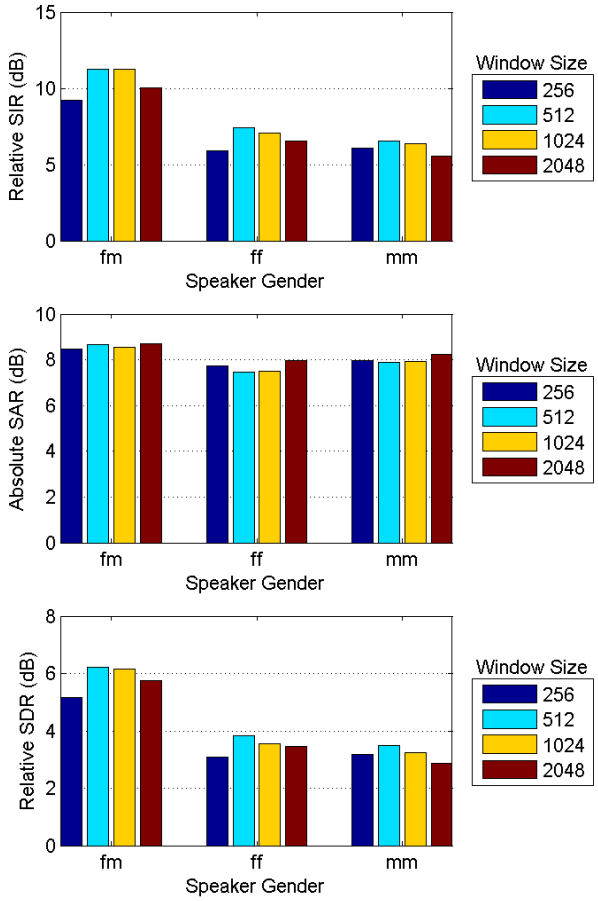

Fig. 3. BSS Eval results illustrating how window size and speaker gender affect performance (averaged over all experiments with β of 0, magnitude exponent of 1, and TMR of 0).

previous paragraph, which are a β of 0 and a magnitude exponent of 1. We see that within each TMR, a window size of 512 performs the best for both the relative SIR and SDR. The absolute SAR is a slightly higher with window sizes of 256 and 2048. What is more interesting in this figure is observing how the BSS Eval measurements are affected by the target speaker's TMR. As the TMR increases, the resulting relative SIR decreases. This can be explained as follows: if the TMR is very low, it is easier to identify and attenuate the masker, but when the TMR is very high, is it more difficult to identify and attenuate the masker. For example, if the TMR were 100 dB, it could be very difficult to find the noise, and attenuating the signal would most likely attenuate the desired signal, thus decreasing the relative SIR. The absolute SAR results can be explained as follows: since the SAR measures the ratio of the signal to the artifacts, if the signal level increases and artifacts stay the same level, then the SAR will improve. However, we see that artifacts do not stay constant over the TMR range, so that for every difference of 5 dB, the artifact level changes by about 2.5 dB. We thus see that an TMR increase of 5 results roughly in an increase of 2.5 dB absolute SAR. And finally, when analyzing the SDR results, we see that as TMR increases, relative SDR decreases. So at +10 dB, although we are able to still increase relative SIR, the artifacts cause just a small improvement in relative SDR. Although the relative SDR and SIR decrease as TMR increases, we fortunately see that the absolute SDR (which is the sum of the mixture TMR and the relative SDR) and absolute SIR increase as TMR increases.

3.2.3. Speaker gender and window size

In Figure 3, we see how the speakers' genders and the window size affect performance. The results plotted are with the optimal parameters discussed in the previous paragraph, which are a β of 0 and a magnitude exponent of 1. We wanted to compare these two variables on the same figure because we had originally hypothesized that the optimal window size would depend on the spectral similarity and the pitch ranges of the speakers. We hypothesized that as spectral similarity increased (same gender), or the pitch ranges decreased (both male speakers), that a longer window would perform better. We found, however, that this was not true. In fact, the 512-point window gives the best overall performance for all three speaker gender combinations (mixed gender, both female, and both male). This is good news because the optimal value of this parameter is not dependent on the speakers' characteristics.

In comparing performance between different sets of speaker genders, we saw that the mixed-gender sentences typically had higher relative SIR, absolute SAR, and relative SDR scores than sentences with the same gender. This is consistent with our hypothesis that more spectrally different sources would result in better separation. We also see that the relative SIR and SDR are higher for the two-female mixtures than the two-male mixtures, and the absolute SAR is just slightly lower. We think that the relative SIR and SDR improvements are higher for the female speakers because there will typically be less overlap in the spectra of the two females because their pitch, and thus spacing between harmonics, is greater. Spectral similarity and spectro-temporal overlap are thus two of the best indicators of NMF's source separation performance. NMF will perform better when similarity and overlap are lower and worse when similarity and overlap are higher.

Here is a summary of the single-channel 2-talker speech separation results:

- The best overall parameters found for relative SDR were a β of 0, a magnitude exponent of 1, and a window size of 512 points (32 ms).
- The Itakura-Saito divergence ($\beta = 0$) with a magnitude exponent of 1.5 and 2 typically maximize suppression of interfering speakers.
- Increasing magnitude exponent typically increases artifacts.
- Increasing TMR typically increases absolute SIR, absolute SAR, and absolute SDR, and decreases relative SIR, and relative SDR.
- A 512-point (32 ms) window works best overall for all speaker gender combinations.
- Performance is negatively correlated with spectral similarity and spectro-temporal overlap of the sources, which give the following order of performance, from best to worst: mixed gender, both females, and both males.

4. MUSIC INTERPOLATION

4.1. Theory

When parts of an audio signal are missing or too corrupted by noise to be recovered by traditional means, spectrogram interpolation can

be used to fill in these gaps [2]. NMF-based spectrogram interpolation works by learning the spectral basis vectors and weights of the observed signal. The missing time-frequency points are ignored in the cost function and update equations. After the algorithm converges, the missing data are replaced by the NMF estimation of those points (see Figure 4). Some methods for estimating the phase of the interpolated data include using random phase, using the phase of a nearby neighbor, and finding a phase that increases STFT consistency [17]. Instead of estimating the missing phase, we just used the phase of the original, uncorrupted signal, so we could focus on the the NMF-based STFT magnitude estimation in our experiments and analysis.

The challenge again is to find the optimal parameters for interpolation. In this task, we will again see how cost function, magnitude exponent, and window size affect performance. Additionally, we will also see how the number of basis vectors affects performance. In the speech separation experiments, we used the training data as spectral basis vectors, but in our interpolation experiments, we have no training data. If we did have training data, such as non-corrupted corrupted data from a different part of the song, we could try using that as training, but the experiments in this paper will use the conditions where training data is available. We will be using absolute SNR of the reconstructed signal to measure performance,

$$SNR = 10 \log_{10} \frac{\| v \|^2}{\| v - \hat{v} \|^2} \tag{9}$$

where v and $\hat{v}$ are the time-domain signals of the orignal and interpolated signals, respectively.

4.2. Experiments and Results

For our experiments, we chose 10-second instrumental excerpts from 4 songs from different genres. They retained the same 44.1 kHz sampling rate as the CD's from which they were extracted. The parameter choices we tested were β (0, 0.5, 1, 1.5, 2), magnitude exponent (0.5, 1, 1.5, 2), window size (1024, 2048, 4096 points, which correspond to 23.2 ms, 46.4 ms, and 92.8 ms, respectively), number of basis vectors (4, 8, 16, 32, 64), and amount of missing data (10-80% at 10% increments). The binary mask for determining the missing data was randomly generated, but each experiment at a given amount of missing data used the same mask to remove additional variance in performance.

The scenarios that we will be focusing on will be when more than 50% of the data is missing. The optimal parameters for when 80% of the data was missing was a β of 1, magnitude exponent of 0.5, window size of 4096 points, and a basis vector rank of 4. We will first discuss how β and window size affect performance, and then we will discuss how the magnitude power and number of spectral basis vectors affect interpolation performance.

In analyzing β, we saw that with a small amount of missing data, β values of 1 and higher performed well, without one being a clear winner, but when more data was missing, a β value of 1 was best. Our theory for this is that since a β value of 0 makes fitting all observed data to be equally important, this may cause undesired behavior when nearly silent (low magnitude) regions exist. As β increases, it puts increasingly more importance on fitting higher energy points. And then as the amount of data increases, the KL divergence performs best perhaps because has a good balance of how it prioritizes fitting higher and lower energy data.

When analyzing window size, we saw that although performance did not dramatically change between different window sizes, a 4096-point window performed best overall. A 4096-point window

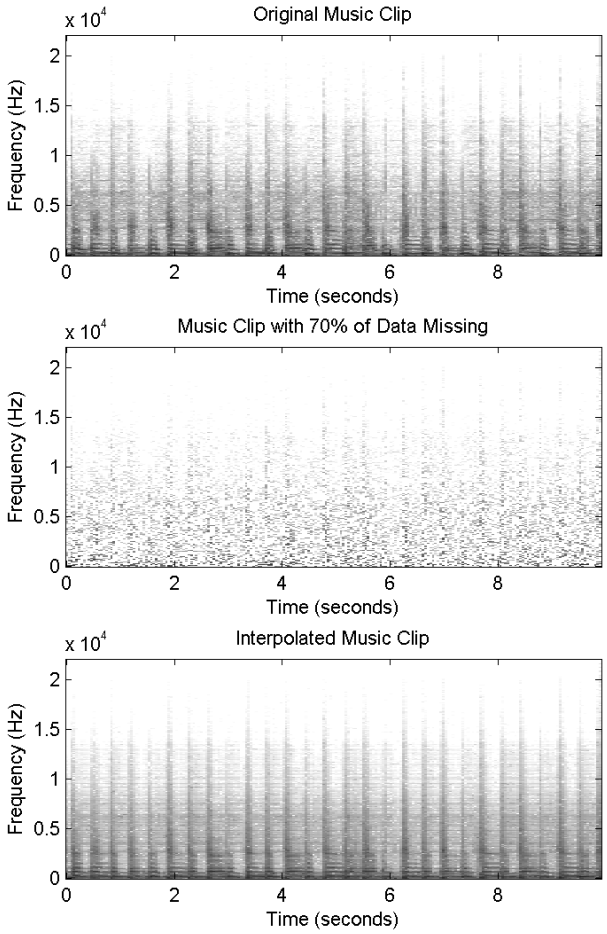

Fig. 4. Example of interpolation of missing music data

at 44.1 kHz corresponds to a 92.8-millisecond window, which is much longer than the optimal-length window for single-channel speech separation (32 milliseconds with 512-point window at 16 kHz sampling rate). This may be due to the differences in temporal rate of music and speech, as well as the difference in tasks. It would be interesting to explore this hypothesis further by comparing the optimal window parameters of interpolating speech, but is out of the scope of this paper.

Next, we will discuss how the magnitude exponent affects performance. In Figure 5, we see that as the amount of missing data increases, the optimal magnitude exponent value decreases monotonically from 1.5 to 0.5. Our observation was that as the missing data increased, the frequency that the estimated data has a much higher energy than is correct is also increased. This seems to happen more with higher values of the magnitude exponent. This is seen most clearly when 80% of the data is missing, where a magnitude exponent of 0.5 performs much better than higher values.

Next, we will discuss how the number of basis vectors affects performance. In Figure 6, we see that as the amount of missing data increases, the optimal number of basis vectors decreases monotonically from 32 to 4. With small amounts of data missing, a higher number of basis vectors, perhaps corresponding to the number of instruments or notes in the sample, performs well. But as the amount of missing data increases, over-fitting becomes an issue, and we

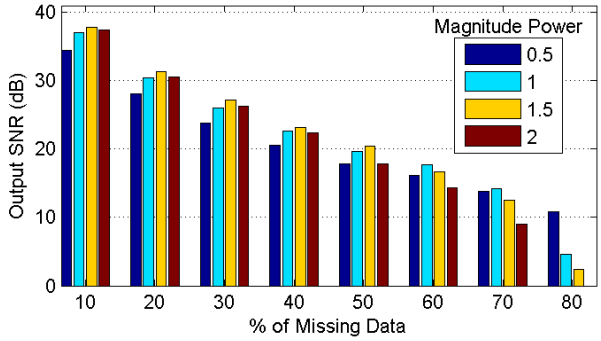

Fig. 5. SNR results illustrating how magnitude power and amount of missing data affect performance (averaged over all experiments with 4 spectral basis vectors, 4096-point window, and β of 1). Missing values indicate negative SNR results, as minimal range value is 0.

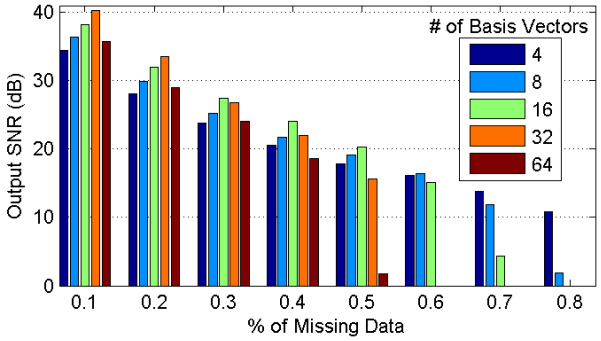

Fig. 6. SNR results illustrating how the number of spectral basis vectors and amount of missing data affect performance (averaged over all experiments with magnitude power of 0.5, 4096-point window, and β of 1) Missing values indicate negative SNR results, as minimal range value is 0.

again see that data estimates being much higher than the correct value become more common with higher values of basis vectors.

Here is a summary of the music interpolation results:

- The best overall parameters, with respect to maximizing the SNR, for interpolating 80% of missing data were a β of 1, a magnitude exponent of 0.5, a window size of 4096 points, and a basis vector count of 4.
- As the amount of missing data increased, the optimal value of the magnitude power and the number of basis vectors tend to decrease.
- As the number of basis vectors increase, the variance of results dependent on the initial values of the basis and weight matrices tends to decrease.

5. CONCLUSION

In this paper, we have seen how parameters and test conditions affect performance for NMF-based single-channel speech separation and music interpolation. Specifically, we analyzed the β, magnitude exponent, and window size parameters and the TMR's and genders of the sources for speech separation, and the β, magnitude exponent, number of basis vectors, and window size parameters and the amount of missing data for music interpolation. We hope that our goals of explaining how parameters affect performance will not only provide optimal parameter choices for these applications, but that the explanations of the data will also provide a deeper understanding and a better intuition of NMF parameters for other applications.

6. REFERENCES

[1] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE TASLP*, vol. 15, no. 1, pp. 1–12, 2007.

[2] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *IEEE MLSP*, 2009.

[3] D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in *INTERSPEECH*, 2005.

[4] P. Smaragdis and J.C. Brown, "Non-Negative matrix factorization for polyphonic music transcription," in *IEEE WASPAA*, 2003.

[5] B. Raj, R. Singh, and P. Smaragdis, "Recognizing speech from simultaneous speakers," in *INTERSPEECH*, 2005.

[6] B. King, P. Smaragdis, and G.J. Mysore, "Noise-Robust dynamic time warping using PLCA features," in *IEEE ICASSP*, 2012.

[7] D. Fitzgerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *IET Irish Signals and Systems Conference*, 2009.

[8] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE TASLP*, vol. 18, no. 3, pp. 528 –537, 2010.

[9] T.O. Virtanen, "Monaural sound source separation by perceptually weighted Non-Negative matrix factorization," *Technical Report*, 2007.

[10] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[11] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[12] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 13, no. 3, pp. 1–24, 2010.

[13] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *NIPS*, 2009.

[14] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *IEEE ICASSP*, 2010.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus*, NIST, 1993.

[17] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *SAPA*, 2008.