SINGLE-CHANNEL SOURCE SEPARATION USING SIMPLIFIED-TRAINING COMPLEX MATRIX FACTORIZATION

Brian King and Les Atlas

Department of Electrical Engineering, University of Washington, Seattle, WA, 98195-2500, USA

ABSTRACT

Although the task seems trivial for human listeners, research in automating source separation still lags far behind human performance and is especially difficult for single-channel signals. One of the latest and most promising methods of single-channel source separation is non-negative matrix factorization, which works by synthesizing signals from a learned set of bases for each source. In this paper, we present a new method of creating these learned sets of bases used in the matrix factorization technique for single-channel source separation. This new method does not suffer the complication of choosing an optimal number of bases as in previous methods. In addition, this paper further explores the new method of complex matrix factorization and compares its performance to non-negative, real matrix factorization for automatic speech recognition of two-talker mixtures.

Index Terms— Non-negative matrix factorization, complex matrix factorization, source separation, speech processing

1. INTRODUCTION

Automatic source separation has been a topic for research for many years, not only because of its difficulty but also because of several compelling potential applications, such as speech enhancement, remixing recordings where individual tracks are unavailable, polyphonic music transcription, and speech recognition by both humans and computers. In this paper, the application most discussed will be automatic speech recognition (ASR), but the algorithm proposed holds potential for use in many other applications as well.

State-of-the-art ASR systems have progressed to the point where they can perform well with a single talker, but rapidly degrade in the presence of multiple talkers or noise [1]. Research in multiple-talker ASR has split in the following two directions: developing a recognizer that works directly with multiple talkers [2,3], and developing a source separation system to preprocess the signal before the ASR system [1]. The latter is the topic of our research presented in this paper.

Many source separation algorithms have been developed, including computational auditory scene analysis [4], independent component analysis [5], blind source separation [6], and machine learning [7]. One method originally applied to image processing that has recently become popular is non-negative matrix factorization (NMF). NMF performs well in some applications, but has some fundamental limitations due to its non-negativity constraints [8,9]. Recent research has extended NMF to allow for complex values, and is called “complex non-negative matrix factorization” [8]. To help avoid confusion, this method will be called “complex matrix factorization” (CMF) in this paper.

This paper starts with a sequence of two introductory sections. The first defines and discusses the differences between NMF and CMF (Section 2). The second introduces a new method of creating bases for NMF or CMF algorithms and discusses its advantages over previous methods (Section 3). Next, automatic speech recognition results on two-talker mixed speech using the proposed new method of creating the bases will be presented and compared with a current state-of-the-art method (Section 4). Performance will be compared on both NMF and CMF versions. The paper will then conclude with a summary and suggestions for future work (Section 5).

2. MATRIX FACTORIZATION

This section will give a background of non-negative matrix factorization (NMF), discuss its limitations, and contrast it with complex matrix factorization (CMF).

2.1. Nonnegative Matrix Factorization

NMF was originally developed and applied to image processing by Lee and Seung [10,11]. The algorithm has since been extended for use on a magnitude spectrum for audio applications, including musical transcription [12], pitch tracking[13], and source separation [14-16]. The NMF algorithm uses an iterative process (such as minimized subgradient descent) to minimize the following cost function,

$$\frac{1}{2} \sum_{n,t} \left| X_{n,t} - \sum_k B_{n,k} W_{k,t} \right|^2 + \lambda \sum_k \left| W_{k,t} \right|^2$$  \hspace{1cm} (1)$$

where $X$ is the short-time Fourier transform (STFT) with usual sampled variables in frequency $\omega$ and time $t$, $B$ and $W$ are the base and weight matrices, respectively, and $\lambda$ controls the influence of sparsity on the weights. The sparsity condition was not part of the original work by Lee and Seung but was added later to allow control over the importance of a smaller mean squared error (MSE) and a sparser weight matrix [17]. In acoustic processing, the columns of $B$ are frequency-domain bases and correspond to an indexed (by $k$) set of bases. In the weight matrix $W_{k,t}$, a column corresponds to the weights given to the bases to reconstruct the frequency content at specified time index $t$ of the original signal. The reason why the algorithm is called non-
negative matrix factorization is because all elements of the matrices are non-negative, meaning

$$|X_{n,j}| \geq 0, B_{n,j} \geq 0, W_{k,j} \geq 0$$  \hspace{1cm} (2)$$

NMF works well in certain applications, but is problematic when used on signals that are overlapping in time and frequency [8,9]. The problem is caused by the fact that although signals may be additive in the complex STFT domain, their magnitudes aren’t additive unless their phases are identical, which can be shown as following:

$$X_{n,j} = C_{n,j} + D_{n,j}$$  \hspace{1cm} (3)$$

$$|X_{n,j}| e^{j\phi} = |C_{n,j}| e^{j\phi} + |D_{n,j}| e^{j\phi}$$  \hspace{1cm} (4)$$

$$|X_{n,j}| = |C_{n,j}| + |D_{n,j}|, \text{iff } \phi = \phi_i = \phi_j$$ $$|X_{n,j}| \neq |C_{n,j}| + |D_{n,j}|, \text{otherwise}$$  \hspace{1cm} (5)$$

where $C_{n,j}$ and $D_{n,j}$ are the two sources in mixture $X_{n,j}$ at a specific time-frequency point. Another, related problem is in recovering phase for the separated signals. Since magnitude-only spectrum matrices are created in the NMF separation step, no phase information is present for synthesizing into a time-domain audible signal. Usually the original mixed phase is used [14], but doing so often causes an audible artifact in the reconstructed signal. Fortunately, the complex matrix factorization presented in the following section avoids these problems by implementing matrix factorization on the complex STFT.

2.2. Complex Matrix Factorization

In order to solve the superposition and phase problems present in the NMF, Kameoka et al. generalized the algorithm for use with complex signals [8]. The weights and bases are still non-negative, but multiplying each time-frequency point is a phase term that allows each spectral base to assume the phase to best fit the mixed signal best:

$$\frac{1}{2} \sum_{n=0}^{N} |X_{n,j} - \sum_{k} B_{n,k} Y_{k,j} e^{j\phi_{n,k}}|^2 + \lambda \sum_{n} W_{n,j} \geq 0$$ (6)

$$B_{n,j} \geq 0, W_{k,j} \geq 0$$ (7)

3. SOURCE SEPARATION USING MATRIX FACTORIZATION

In this section, we describe how to use matrix factorization for source separation, summarizing past methods and then presenting our novel method. Source separation works because of the summation of the bases embedded in the matrix multiplication. In order to separate the signals after the base $B$ and weight $W$ matrices are calculated, the basis elements and weights corresponding to a specific source are multiplied together (see separation step in Figure 1).

In using matrix factorization for source separation, there have been two main methods for calculating the bases [8]. The first, which we will call the “no-train” method, calculates the base and weight matrices from the mixed signal without a training step. There are two problems with this method. The first is that the order of the bases is random so that it is unknown which basis vectors correspond to which source. Because of this, basis elements have to be further analyzed manually or otherwise to try to determine which source they represent. The other problem with this method is that the number of basis vectors must be specified before factorization. The difficulty in choosing the number of vectors is that if the count is too small, the basis set consists of spectra that contain multiple speakers or phonemes. If the count is too high, however, the set of frequency vectors in the set may be too general and not speech-like. For example, if the base matrix is overcomplete, one possible basis decomposition would be the identity matrix, which obviously is not useful for multi-talker separation. Thus, picking an appropriate number of basis vectors is usually done in an ad hoc manner by time-consuming trial-and-error, which highlights the need for a better solution.

The second previous method, which we will call the “factorize-to-train” method, for calculating bases solves the first problem mentioned above by first calculating the base matrix with clean, single-source training data. If there are $N$ sources in the mixture, then a base matrix is calculated from each clean source during the training. All these bases are concatenated together in order to determine the weight matrix for the mixed signal. Once the weight matrix is calculated, the separation step is possible because the basis indices are known for each source. However, this method still suffers from the latter problem mentioned above, in that the base rank must still be chosen.

Our new, proposed method for choosing the base, which we will call the “copy-to-train” method, does not have either of the problems mentioned above, and can be seen in the above block diagram in Figure 1. Instead of calculating the base matrices for each source by matrix factorization, the base is simply set to be the magnitude of each single-source training STFT. In other words, the bases are a “copy” (in STFT form) of the original
training data. This approach is advantageous in these ways: (1) separation becomes easy as the base indices are known for each source, (2) the problem in choosing the number of bases is eliminated, (3) the frequency vectors of the bases are known to correspond to the sound units of the speaker, and (4) processing the bases is significantly faster since no matrix factorization is needed in training.

4. EXPERIMENTAL DESIGN AND RESULTS

In order to test our new copy-to-train method, we designed a pilot study to compare it with the factorize-to-train. The experimental data consisted of hand-transcribed English speech from one male and one female speaker, which were extracted from *Meet the Press*, a television broadcast news show. The training data consisted of one minute of speech from each. The mixed signal was created by taking clean speech from each talker and mixing them together at several (-6, -3, 0, +3, +6 dB) target-to-masker (TMR) ratios. The clips used for the mixed speech were from the same speakers, but different segments of speech so that there was no overlap in the training and mixed signals. The mixed signal was eight seconds long and consisted of 59 words, 28 from one speaker and 31 from the other. The mixed signals were then processed the following four ways: NMF and CMF with the factorize-to-train method, and NMF and CMF with our new copy-to-train method. The separated signals were then processed by the ASR system [18] and scored against their correct transcriptions. The mixed signals were scored against both speakers’ transcriptions. All ASR parameters were fixed for all the tests so that changes in ASR results would be from the separation steps.

With this experiment, we made three primary comparisons. The first was to compare the performance of the factorize-to-train and copy-to-train methods. The all-in-one method that learns the bases from the mixed signal is not included in the comparison because there is not an accurate way yet to determine which of the bases correspond with which speaker. Figure 2 graphs the percentage improvement of the new method over the previous method over the entire range of TMR’s used using the following equation:

\[
\frac{C_{new} - C_{prev}}{C_{prev}}, \quad C \text{ is average } \% \text{ correct for all TMR's}
\]  

(8)

The previous factorize-to-train method was used to calculate 50, 100, and 200 bases for each speaker. These base counts were chosen because they are typical values that have been used previously and are not meant to be an exhaustive list. Looking first at the NMF comparison, the new copy-to-train method performs better than with 50-base set, but performs worse than the 100-base and 200-base sets. The new method using CMF, on the other hand, outperforms the previous method in all three cases. We hypothesize that the CMF performs better because the complex model is a better representation by satisfying superposition and learning phase.

The next comparison, which is the most important in terms of practical needs, was to see how the CMF separation performed versus the unprocessed signal, using the copy-to-train method. Figure 3 graphs the percentage correct for each TMR as well as the original speech signals before they were mixed. The CMF separation outperformed the unprocessed mixture in every instance, except for the original, unmixed signal. In that case, the degradation was slight, but in such cases it would be easy to detect the presence of clean speech so that the processing could be bypassed in such instances.

The third comparison was between the performance of the CMF and NMF, using the copy-to-train method. Figure 4 graphs the percentage improvement of the CMF over the NMF using the copy-to-train method. The results show that the CMF performs better than the NMF in every instance. We think that the degraded performance may be due to the audible artifacts present in the NMF results. The artifacts sound like a stationary noise in the higher frequencies with an envelope corresponding to the envelope of the separated signal. These artifacts are likely caused by the superposition and phase problems present in the NMF technique. In contrast, the CMF has no audible artifacts. In fact, a good analogy for the CMF result is that it sounds as if the masker was attenuated in the total mix, as if turned down by a fader on a mixing console. The fact that there are such insignificant artifacts is rare in the area of source separation and speech enhancement, because the majority of such processes leave significant artificial-
Fig. 4. ASR Percent improvement of CMF over NMF using new method of training bases

sounding artifacts which constrain their use to situations where sound fidelity is less important.

This pilot study has shown promising initial results for the new copy-to-train method, though in order to gather meaningful performance statistics, a larger study with more training and test data is needed and is underway.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the copy-to-train method, a new method for talker separation that has many advantages over previous methods. We further showed in preliminary experiments that this method may outperform the previous factorize-to-train method for preprocessing mixed speech for ASR, though more experiments are necessary. We also contrasted CMF with NMF, and showed that it may perform better than NMF with the copy-to-train method.

There are many directions for future research. The first will be to conduct similar experiments on a larger dataset to gain a better understanding of the new method’s significance, predicting performance based on the characteristics of the training and mixed sets, like number of sources, different types of sources (such as speech mixed with noise or music) length of training data, acoustic similarity between sources, and TMR. Another direction is looking into how performance is affected if the system is trained on fewer than all the sources, such as just the target or just the maskers. An interesting area of research is related to the sparsity of the weight matrix, and whether we can take advantage of the sparsity to increase computational speed and separation performance.

This research was funded by a grant from the Air Force Office of Scientific Research and a gift from Adobe Systems Incorporated. Special thanks to Elliot Saba, Mari Ostendorf, Alex Marin, and Brian Hutchinson for ASR and computing support.

6. REFERENCES


