BAYESIAN TWO SOURCE MODELING FOR SEPARATION OF N SOURCES FROM STEREO SIGNALS

Aaron S. Master

Stanford University
Center for Computer Research in Music and Acoustics
Stanford, California 94305-8180, USA

ABSTRACT

We consider an enhancement to the DUET sound source separation system [1], which allowed for the separation of N localized sparse sources given stereo mixture signals. Specifically, we expand the system and the related delay and scale subtraction scoring (DASSS) [2] to consider cases when two sources, rather than one, are active at the same point in STFT time-frequency space. We begin with a review of the DUET system and its sparsity and independence assumptions. We then consider how the DUET system and DASSS respond when faced with two active sources, and use this information in a Bayesian context to score the probability that two particular sources are active. We conclude with a musical example illustrating the benefit of our approach.

1. INTRODUCTION

Sound source separation refers to the problem of synthesizing N source signals given an M channel mixture of those source signals. When there are fewer input mixtures than sources to be separated ($M < N$), we have the *degenerate* case. In the degenerate case, it is necessary to use prior information about the source signals to perform demixing, because of the ill-posed nature of the inverse mathematical problem.

We presently consider the two mixture degenerate case. In digital audio, we frequently encounter this case, as many or most currently available commercial digital recordings contain two channels (stereo) but more than two instruments, voices, or other sounds.

A variety of approaches to this and other degenerate problems have been tried [3]. Each method exploits one or more features of the sound sources, as they must do in order to be successful. Such features include the sources' time-frequency sparsity, their time-frequency independence, and their distinct amplitude and delay characteristics between the mixtures. A brief review of these techniques for the two source case is included in [1].

We find that the DUET system [4, 5, 1] has achieved particularly convincing results, but can still be improved. Specifically, we note that the system only works as intended when in fact the sources are distinct in time-frequency space. This is referred to as "source sparsity" although non-overlap of sources is also required. This is because co-occurring sparse sources cannot be separated. In performance of tonal Western music, sources are in general sparse because instrumental ranges are finite and most compositions do not require constant playing or singing throughout time. The sources, however, are not in general independent, unless the ensemble is without skill or the music requires that players sound notes in a deliberately random fashion. The harmonic nature of Western music exacerbates the problem, because harmonics whose fundamental frequencies are in (possibly imperfectly) consonant relations will overlap. Even in the case of dissonant or deliberately random music, pitches are in general discretized to the 12-tone Western scale, leading to overlap of some harmonics.

Given these facts, it is necessary that the DUET system be modified if it is to deal with non-independent sources such as those seen in music. Presently, we consider a method for the case when exactly two unknown sources are present. This means that two instruments or voices are sounding though we do not know a priori if it is, for example, the bass and cello or cello and flute. Clearly, this case is only an incremental improvement of the current one-source-at-a-time system. However, in the cases of musical trios or four speaker examples, the two-source assumption is of great benefit.

To consider the benefit in the current approach, we first review the DUET system and the related delay and scale subtraction scoring (DASSS) [2], and explore how these models are affected when two sources are present at the same point in time-frequency space. In the third section, we consider how to exploit the two-source system response in a Bayesian context. Specifically, we develop a method for scoring the probability that two particular sources are active given DASSS data. We conclude with a musical example showing the efficacy of using Bayesian Modeling of DASSS data rather than DUET for determining and demixing two active sources.

2. DUET AND DASSS REVIEW

We first review the DUET system [4, 5, 1] of Scott Rickard and other authors. The DUET system performs sound source separation of N sources from two channels, where N is in general greater than two. The DUET system assumes the following STFT domain linear mixing model for sources S_i in left channel X_1 and right channel X_2:

$$X_1 = S_1 + S_2 + \cdots + S_N \tag{1}$$
$$X_2 = a_1 e^{-j\omega\delta_1} S_1 + a_2 e^{-j\omega\delta_2} S_2 + \cdots + a_N e^{-j\omega\delta_N} S_N \tag{2}$$

where a_i represents the scale parameter and δ_i represents the delay parameter, each from the left to right channel, for some source i. We refer to a_i and δ_i together as the *mixing parameters* for a given source i.

By assuming that only one source at a time is active in time-frequency space – a near-realistic assumption for independent speech

0-7803-8484-9/04/$20.00 ©2004 IEEE

sources – we may estimate the mixing parameters for a particular time-frequency point via:

$$(a_i, \delta_i) = \left(\frac{|X_2(\omega_k, \tau)|}{|X_1(\omega_k, \tau)|}, \Im \left\{ \log \left(\frac{X_1(\omega_k, \tau)}{X_2(\omega_k, \tau)} \right) \right\} / \omega_k \right). \quad (3)$$

After collecting many such estimates, the DUET system prepares a two-dimensional histogram whose peaks in (a_i, δ_i) space should reveal the mixing parameters for each of the N sources. To demix the sources, DUET considers the set of parameter estimates a second time after the source mixing parameters are estimated from the histogram. It then assigns each point in time-frequency space to the source whose mixing parameters are closest to that estimated for the time-frequency point. To do this, a variety of matching schemes may be used. We have presented delay and scale subtraction scoring (DASSS) [2], which is similar to a method presented recently by the original DUET authors in [1].

In DASSS, we define a set of functions Y_i such that:

$$Y_i \quad \equiv \quad X_1 - \frac{1}{a_i} e^{+j\omega\delta_i} X_2 \quad (4)$$

and the mixing parameters are always treated as known quantities. If in fact exactly one source, S_g, is active at a given frequency bin in a given frame, it may be shown that our model predicts:

$$\begin{aligned} \hat{Y}_{i=g} &= 0 && (5) \\ \hat{Y}_{i \neq g} &= \alpha_{j,i} S_j && (6) \\ &= \alpha_{j,i} X_1. && (7) \end{aligned}$$

where

$$\alpha_{u,v} \equiv (1 - \frac{a_v}{a_u} e^{j\omega(\delta_u - \delta_v)}). \quad (8)$$

We now observe that we may similarly predict the DASSS function values Y_i when two sources S_u and S_v are active:

$$\begin{aligned} \hat{Y}_{i=u} &= \alpha_{uv} S_v && (9) \\ \hat{Y}_{i=v} &= \alpha_{vu} S_u && (10) \\ \hat{Y}_{i \neq (u|v)} &= \alpha_{iu} S_u + \alpha_{iv} S_v && (11) \end{aligned}$$

We now make an important observation. If we know how S_v and S_u are distributed, we then know how $\hat{Y}_{i=u}$, $\hat{Y}_{i=v}$, and $\hat{Y}_{i \neq (u|v)}$ are distributed. (In general, we will see that distributions on $|S_u|$ and $|S_v|$ may be practically estimated from knowledge about a musical or speech source, such as its range and loudness. Distributions on $\angle S_u$ and $\angle S_v$ are not informative, and thus we will use the set $|Y_i|$, $i \in \{1...N\}$ rather than the sets Y_i or $\angle Y_i$ as our DASSS data.) Below, we will exploit our knowledge of the DASSS data in a Bayesian context to determine if (and which) two sources are most likely active.

Much as we know how the DASSS data Y_i functions will behave for the two source case, it may also be shown [6] that we can predict the values for the DUET data given by equation 3 in the same case. It is not practical, however, to exploit this data, as logistics and computation quickly become prohibitive [6]. We therefore focus our efforts on DASSS data below.

3. BAYESIAN FRAMEWORK

As suggested above, the DASSS data produced by equation 4 may in fact reveal *which* two sources are active at a particular point in time-frequency space if exactly two sources are active. This is very useful information, because once the active sources are known, they may be demixed by solving for S_u and S_v in:

$$\begin{bmatrix} S_u \\ S_v \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ a_u e^{-j\omega\delta_u} & a_v e^{-j\omega\delta_v} \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad (12)$$

which follows directly from equations 1 and 2 when only two sources are active.

Formally, we express the two most likely sources given some DUET or DASSS data D as those maximizing $p(u, v|D)$. Applying Bayes' rule, we can express this as

$$p(u, v|D) \quad = \quad \frac{p(D|u, v)p(u, v)}{p(D)}. \quad (13)$$

We can see by inspection of equations 9 through 11 that the STFT frequency under consideration, ω, affects DASSS data D (the Y_i values). So, we are mindful that the problem is a different one for each of the frequency values ω under consideration.

Since $p(D)$ is not a function of u or v, it is possible to discard it from the maximization (though we may wish to use it later if a confidence measure is sought). The quantity $p(u, v)$ is largely estimated with musical knowledge. For example we may know that the clarinet (source $u = 1$ for example) tends to play at the same time as and less loudly than the violin (say, source $v = 2$), whose frequency components tend to be at harmonics of frequencies above 200 Hz, and rarely throughout time. Though very useful, such information is not within our current signal processing interest, and is not considered now. (For now, we will treat all $p(u, v)$ as equally likely.)

We are left, then, to consider $p(D|u, v)$ for each ω, the probability that particular DASSS data is produced when sources v and u (but no others) are active at frequency ω. To this end, we next explicitly return to the distributions suggested by equations 9 through 11. By doing so, we identify the necessary values of $p(D|u, v)$ where D represents DASS scores Y_i.

4. BAYESIAN FRAMEWORK APPLICATION

As mentioned above, knowledge of the distributions on $|S_u|$ and $|S_v|$ can be used to create distributions on the $|Y_i|$ values from equations 9 through 11. The distributions on $|S_i|$ need not be Gaussian to use the technique described here. Assuming this simplifies the situation, however, and informal experiments have shown that we may fairly model the amplitude of the STFT coefficients this way by considering their distribution as the positive values of a zero-mean Gaussian. Specifically, it may be shown that doing so leads to

$$\begin{aligned} |\hat{Y}_{i=u}| &\sim N(0, \sigma_v^2 \cdot |\alpha_{uv}|^2) && (14) \\ |\hat{Y}_{i=v}| &\sim N(0, \sigma_u^2 \cdot |\alpha_{vu}|^2) && (15) \\ |\hat{Y}_{i \neq (u|v)}| &\sim N(0, \sigma_u^2 \cdot |\alpha_{iu}|^2 + \sigma_v^2 \cdot |\alpha_{iv}|^2) && (16) \end{aligned}$$

where σ_i^2 represents the variance for source i at a given frequency. (We again recall that these distributions, and their related σ^2 and α values are source dependent, and different for each frequency.

Clearly, σ^2 will be larger for frequencies corresponding to the active range of a given voice or instrument.)

We may calculate then, the probability that a set of data D (in the form of $|Y_i|$ values given by equation 4) was generated by the presence of sources S_u and S_v via:

$$p(D|u,v) = \prod_{i=1}^{N} p(|Y_i||u,v) \qquad (17)$$

where

$$p(|Y_i||u,v) = \frac{1}{\sqrt{2\pi \mathrm{var}(\hat{Y}_i|u,v)}} \exp\left[\frac{-|Y_i|}{2\mathrm{var}(\hat{Y}_i|u,v)}\right]$$

and $\mathrm{var}(\hat{Y}_i|u,v)$ refers to the variance in the distributions of expressions 14 through 16.

To achieve our goal of determining the two most likely sources at a given point in time-frequency space, we first determine $p(D|u,v)$ for the point's $|Y_i|$ values using equation 17 and considering every possible (u,v) combination. Then we substitute in our result to equation 13 which allows us to take into account prior probabilities. By allowing u or v to be "NULL" and assigning a value corresponding to the noise floor as the variance of S_{NULL}, we effectively can include the one-source combinations used in the DUET system as well.

5. A PATHOLOGICAL MUSICAL EXAMPLE

We have applied the current Bayesian Two Source Modeling (BTSM) technique to a pathological musical example that is otherwise particularly difficult to deal with. We consider a musical trio in which two sources are always active, and each plays the same note in the same octave. The samples are of clarinet, violin, and cello, and come from the Iowa samples database [7].

As can be seen from the spectrograms in figure 1, the clarinet and violin first play together, then the clarinet and cello, and finally the violin and cello. The mixing was done synthetically as specified by the DUET signal model, with mixing parameters:

source	a_i	δ_i
1	1.05	-9.07e-5
2	1.01	-2.27e-5
3	0.9	6.80e-5

To prepare the system, we first processed excerpts (segmentation courtesy of Pamornpol 'Tak' Jinachitra) from all of the files for each instrument to gain estimations of the variance of each source's STFT magnitude coefficients. We then processed all points in STFT space for the test file containing X_1 and X_2, calculating $p(u,v|D)$ using the Bayesian approach above, and including null sources to allow a one source output. We used a uniform prior $p(u,v)$, indicating no preference for the activity of any one or two of the three sources.

In the spectrograms in figure 2 and the output SNRs (dB) in the table below, we see the results achieved by the DUET system and the current BTSM system. Though the DUET system often does separate some of the frequency components correctly, its single active source constraint becomes a liability when most frequency components of the sources overlap. Indeed, we can see cases in the figure where components sharply enter or exit, a highly audible phenomenon. The BTSM approach achieves much higher SNR, and allows sharing of frequency components between two sources. We see that it sometimes chooses the two active sources incorrectly, giving data to the violin, for example, when only the clarinet and cello are active. More often than not, however the system guesses correctly about which two sources are active, and makes less audible errors. Time domain envelope plots (whose

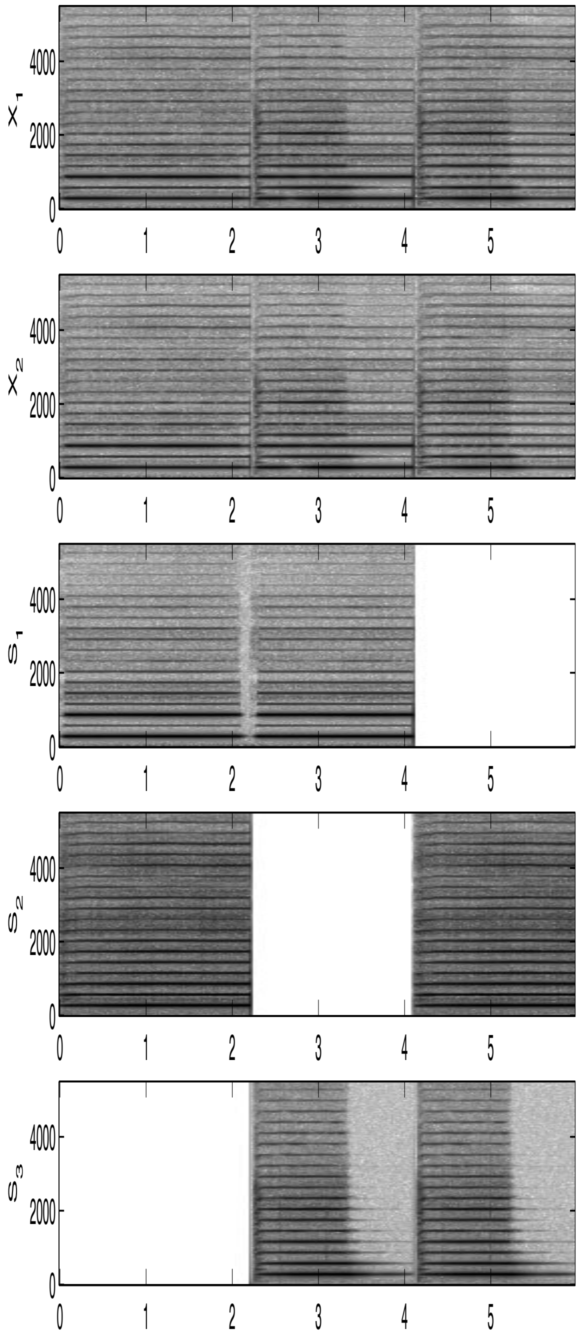

Fig. 1. The original mixtures X_1 and X_2 and the source performances S_1 (clarinet), S_2 (violin), and S_3 (cello) used to make them.

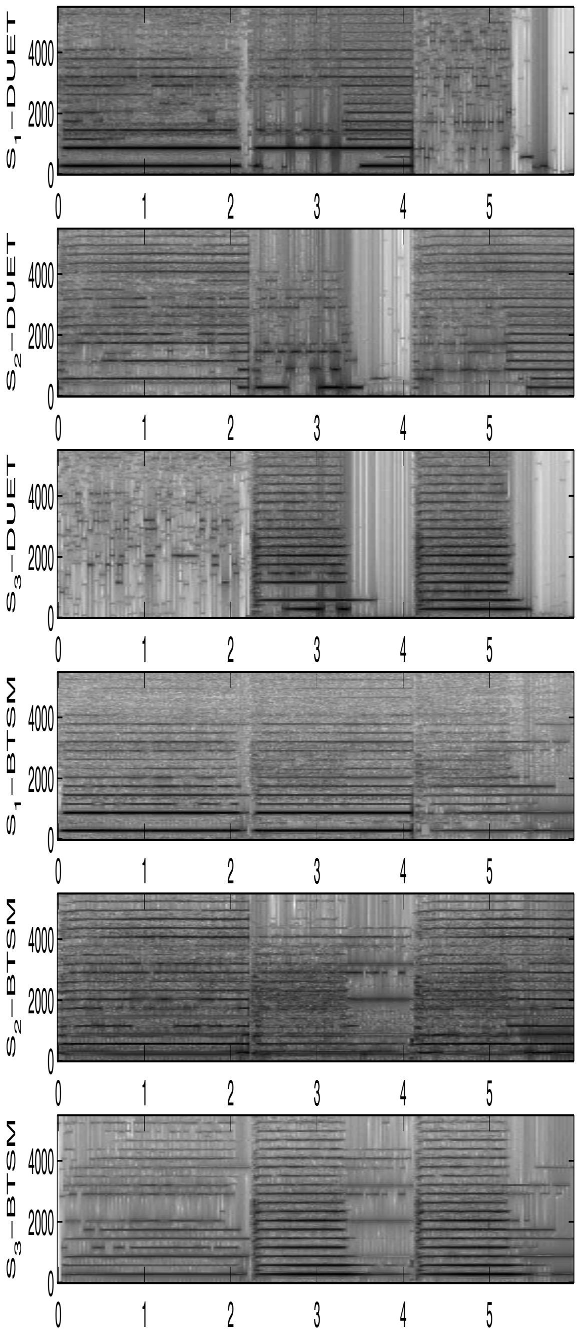

Fig. 2. The separated original mixtures achieved by the previous DUET approach and the new BTSM approach.

inclusion is prevented by space issues) confirm the above.

source	Input SNR	DUET SNR	BTSM SNR
1	-0.4	7.1	**15.7**
2	-13.2	-6.0	**1.1**
3	-0.5	6.3	**18.3**

6. SUMMARY AND FUTURE DIRECTIONS

We have presented a Bayesian framework in which it is possible to estimate the probability that a particular 1 or 2 of N sources are active at a single point in STFT time-frequency space, given a stereo mixture signal and the mixing parameters for the sources. This is significant because it allows us to demix up to two, rather than just one, source for each time-frequency point. This is an important advancement in signals where sources overlap, namely musical signals. Further, the system allows us to bias the probabilities in favor of which sources are more likely to be present at given frequencies, which can be especially beneficial when voices or instruments have known frequency component ranges. Currently, the system does not explicitly identify the probability that three or more sources are active, however.

In the future, we will consider the prior probabilities of sources in nontrivial detail, and seek to integrate data across frames and frequency components. We will also consider an iterative approach for source estimation in which we update priors for the active sources at a given moment based on the data in local frames.

7. REFERENCES

[1] Ozgur Yilmaz and Scott Rickard, “Blind separation of speech mixtures via time-frequency masking,” in *Submitted to IEEE Transactions on Signal Processing, November 4, 2002*.

[2] Aaron S. Master, “Sound source separation of n sources from stereo signals via fitting to n models each lacking one source,” Tech. Rep., CCRMA, Stanford University, 2003, Available from http://www-ccrma.stanford.edu/˜asmaster/.

[3] Andre J. W. van der Kouwe, DeLiang Wang, and Guy J. Brown, “A comparison of auditory and blind separation techniques for speech segregation,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 189–195, March 2001.

[4] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures.,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000, vol. 5, pp. 2985–2988.

[5] Scott Rickard and Ozgur Yilmaz, “On the approximate w-disjoint orthogonality of speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, May 2002, vol. 3, pp. 3049–3052.

[6] Aaron S. Master, “Bayesian two-source models for stereo sound source separation of n sources,” Tech. Rep., CCRMA, Stanford University, 2003, Available from http://www-ccrma.stanford.edu/˜asmaster/.

[7] Lawrence Fritts, *University of Iowa Musical Instrument Samples*, 1997, available online at http://theremin.music.uiowa.edu.