Perception of phase changes in the context of musical audio source separation

Chungeun Kim¹, Emad M. Grais¹, Russell Mason², and Mark D. Plumbley¹

¹Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom
²Institute of Sound Recording, University of Surrey, United Kingdom

Correspondence should be addressed to Mark D. Plumbley (m.plumbley@surrey.ac.uk)

ABSTRACT

This study investigates into the perceptual consequence of phase change in conventional magnitude-based source separation. A listening test was conducted, where the participants compared three different source separation scenarios, each with two phase retrieval cases: phase from the original mix or from the target source. The participants’ responses regarding their similarity to the reference showed that 1) the difference between the mix phase and the perfect target phase was perceivable in the majority of cases with some song-dependent exceptions, and 2) use of the mix phase degraded the perceived quality even in the case of perfect magnitude separation. The findings imply that there is room for perceptual improvement by attempting correct phase reconstruction, in addition to achieving better magnitude-based separation.

1 Introduction

Along with the advancement of the so-called computational auditory scene analysis (CASA) lies the progress of sound source separation, where typically a specific target sound is extracted from a mixture of several sources. Its wide areas of applications include speech enhancement [1], sound event detection along with classification [2], and remixing of audio contents [3]. Various signal processing techniques have been known and used, depending on how the mixed sound was produced, such as the number, spatial distribution or spectral characteristics of the individual sources, and the details of capturing configuration.

One of the conventionally used techniques in the separation of musical mixes is time-frequency masking [4, 5, 6], which is known to be useful in particular when the mixed sound is known to consist of sources with spectro-temporal diversity. In this approach, the mixed signal is converted into and viewed in time-frequency domain via a Short-Time Fourier Transform (STFT). Approximations of various target sources are made as time-frequency masks, by which the spectrogram of the mixed signal is multiplied leading to separation. More specifically, in the case of single-channel source separation from a mixture of I audio sources as y(t) = \[ \sum_{i=1}^{I} s_i(t) \], the aim is to find estimates \( \hat{s}_i(t) \) for the sources \( s_i(t) \), \( i \) from the mixed signal \( y(t) \). This can be formulated in the time-frequency domain as \[ Y(n, f) = \sum_{i=1}^{I} S_i(n, f) \], where \( S_i(n, f) \) is the unknown
STFT of source $s_i(t)$, $Y(n,f)$ is the STFT of the observed mixed signal $y(t)$, $n$, and $f$ are the time and frequency indices respectively. The estimate of the target source $i$ is computed by multiplying the mask of source $i$ with the magnitude spectrogram of the mixed signal:

$$\hat{S}_i(n,f) = M_i(n,f) \times Y(n,f)$$  \hspace{1cm} (1)

where $\hat{S}_i(n,f)$ is the estimate of the magnitude spectrogram of the target source $i$, $Y(n,f)$ is the magnitude spectrogram of the mixed signal, and $M_i(n,f)$ is the spectral mask predicted from the source separation algorithm being applied. The resultant estimate is converted back to the time domain as the final output.

The performance of the separation or the quality of the separated outputs depends on how closely these masks represent the original individual sources (referred to as references, whether or not available) before mixing. The recent emergence of machine learning technology has greatly reinforced the ability to predict the time-frequency masks for known types of sources, with the help of increasing amount of audio data per source category for the training process [7, 8]. This, together with the progress of the machine learning techniques, has led to continuous improvement of the separation performance.

Nevertheless, one non-ideal by-product of this approach with potential perceptual effects is the loss of phase information. In some of the above-described source separation processes, the estimations are made with the magnitude or power spectrograms for the target sources. At the reconstruction stage, the phase of the mixed signal is often used for the conversion into the time domain, instead of the phase information of the individual sources which is unknown. Many studies have considered improving the quality of the separated sources by incorporating the phase information in the separation processes [9, 10, 11, 12]. In [11], for example, a deep learning-based musical audio source separation algorithm, with the addition of the phase information of the target, was found to perform better than the same technique without the phase information. However, the performance comparison was made in terms of three objective energy-based metrics. The recent growing question over the ability of these metrics to accurately predict the actual perception by the listeners [14], performance evaluations have been made conventionally with these physical measures.

Therefore, this study aims to investigate into the perceptual importance of phase preservation of audio sources in the context of source separation, by means of a listening test. It was intended that this study would help to better understand whether the correct phase information in source separation would be beneficial in perception, as well as in terms of the energy-based physical measures.

In the following sections, the details of the experiment are firstly described, together with some technical background of the source separation approaches introduced for this study. Then the results are analyzed and discussed further.

## 2 Perceptual evaluation of variations of phase retrieval

This section outlines the listening test conducted to investigate into the perceptual effects of the perfect phase information in source separation.

### 2.1 Experimental design

For the test, three different cases of mono source separation were considered. This was to introduce a wide range of perceivable performance differences based on the magnitude separation, along with the intentional manipulation of the phase. The following subsections describe more details on the design towards the creation of the stimuli, and on the test method.

#### 2.1.1 Sources for stimuli

The stimuli for the experiment were created from a group of songs from the SiSEC-2015-MUS-task dataset [15]. The dataset has 100 stereo songs with different genres and instrumentations provided with separate individual tracks of the vocals, bass, drums, and the others, which can be used for training, validation or testing of source separation algorithms. For this experiment, we converted the stereo songs into mono by computing the average of the two channels for all songs and sources in the data set. Then with 50 of the songs categorized as the test dataset, we calculated the phase differences between the original mix and the
original vocal track. Among the songs that showed relatively large overall phase differences across time and frequency, firstly those with little vocal activities were excluded as irrelevant. Then through listening to the individual songs and tracks, 15-second clips were extracted from the following three songs in the test dataset:

- Song 45: Johnny Lokke - Promises & Lies (Classic Heavy Rock)
- Song 67: Sambasevam Shanmugam - Kaathaadi (Bollywood)
- Song 91: Traffic Experiment - Sirens (Melodic Alt Rock)

with the intention to introduce as large perceptual differences potentially as possible between the original mix phase and the original vocal phase. The three source separation scenarios were subsequently applied as described in the following subsections.

### 2.1.2 Case 1: source separation in practice

For this scenario, we used the fully convolutional neural network (FCN) as an example for source separation based deep learning model, which has shown better audio source separation performance than the feed forward and recurrent neural networks [16, 17]. The 50 songs from the training dataset were used to train the FCN, which was then applied to the test dataset as described above. The data were sampled at 44.1kHz. The magnitude spectrograms for the data were calculated using the STFT, with a 2048-point Hanning window and a hop size of 512. For more thorough technical background for this source separation method, the readers are encouraged to refer to [17, 18].

For the input and output data for the FCN, we chose the number of spectral frames in each 2D-segment to be 15 frames ($N = 15$). This means the dimension of each input and output instant for the models is 15 (time frames) × 1025 (frequency bins) as in [18]. Thus, each input and output instant (the 2D-segments from the spectrograms) spans around 209 msec of the waveforms of the data.

Table 1 shows the number of layers, the number of filters in each layer, and the size of the filters for the FCN. For example, the first layer in the FCN has a set of 21 filters with size $11 \times 819$ (where 11 is the size of the filter in the time-frame direction and 819 in the frequency direction of the spectrogram). Thus, the first layer generates 21 feature maps. Each feature map is $15 \times 1025$ (the same size of the input and output segments). The activation function in the last FCN layer is the rectified linear unit (ReLU) and in the remaining layers is the exponential linear unit (ELU).

### Table 1: The detailed structures of the FCN.

```
<table>
<thead>
<tr>
<th>Layer number</th>
<th>The input/output data with size 15 frames and 1025 frequency bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conv2D[21,11×819]</td>
</tr>
<tr>
<td>2</td>
<td>Conv2D[32,7×33]</td>
</tr>
<tr>
<td>3</td>
<td>Conv2D[64,3×5]</td>
</tr>
<tr>
<td>4</td>
<td>Conv2DTrans[64,3×5]</td>
</tr>
<tr>
<td>5</td>
<td>Conv2DTrans[32,7×33]</td>
</tr>
<tr>
<td>6</td>
<td>Conv2DTrans[21,11×819]</td>
</tr>
<tr>
<td>7</td>
<td>Conv2DTrans[1,15×1025]</td>
</tr>
</tbody>
</table>
```

The parameters for the FCN were initialized randomly. FCN was trained using back-propagation with gradient descent optimization using Adam [20] with parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, a batch size 100, and a learning rate which started from 0.0001 and was reduced by a factor of 10 when the values of the cost function did not decrease on the validation set for three consecutive epochs. The maximum number of epochs was 25. The algorithm was implemented using Keras with Tensorflow backend [21].
2.1.3 Case 2: unrealistic source separation, closer to ideal case

For this scenario, we introduced the ideal binary mask (IBM) as the source separation technique applied in the process described earlier with Eq. (1). The binary mask is one of the most used masks in source separation [22], which assumes that at each time-frequency bin in the spectrogram of the mixed signal there is only one source active. The IBM is usually used to show the best source separation performance that the binary mask can achieve. It assumes the reference target sources are given, which is an unrealistic situation. However, it shows the upper boundary for the quality which the binary mask can achieve. The IBM for source $i$ is defined as:

$$M_{IBM}(n,f) = \begin{cases} 
1 & \text{if } S_i(n,f) \geq S_j(n,f), \forall j \neq i \\
0 & \text{otherwise}
\end{cases}$$

(2)

where $S_i(n,f)$ is the magnitude spectrogram of the reference target source (the vocal track in this experiment) $i$ at time frame $n$ and frequency bin $f$.

2.1.4 Case 3: ideal source separation, reference magnitude

For this scenario, we assumed that we had a perfect source separation. Therefore, the original target vocal magnitude (the reference magnitude) was used without any processing, only with the phase information to be manipulated.

2.1.5 Phase manipulation

For each of the three source separation scenarios described above, two cases of phase retrieval were assumed. The first case corresponded to having the incorrect phase information, in other words, the phase from the original mix. The second case corresponded to having the perfect phase, or the phase from the original target vocal. Although the second case is not realizable in practice, the intention was to compare the typical phase retrieval approach to the ideal case where the perceptual effects would potentially be the largest. This led to a total of six magnitude-phase combinations per song to be used for the listening test.

2.1.6 Test method and interface

Pairwise comparison was used as the test method, mainly with the intention to detect any perceptual difference between the mix phase and the target vocal phase cases. The unprocessed vocal track was additionally provided as an external reference. The Web Audio Evaluation Tool [23] was used to create the user interface for the test, which ran on a PC with an external USB audio interface (Focusrite Scarlett 2i2). The six magnitude-phase combinations per song led to $6 \times 15 = 90$ pairs to be compared per song, and 45 pairs for all 3 songs. All of the pairs were presented in random order, with the corresponding external references. All stimuli were loudness normalized to -23 LUFS in accordance with the EBU R128 loudness normalization procedure, and were presented over a pair of headphones (AKG K271 Mk II).

2.2 Participants and test procedure

The test procedure was assessed and approved in accordance with the University of Surrey ethics guidelines. A total of 26 subjects participated in the listening test. They were given written and verbal instructions on the experiment. Written consents were then obtained. No hearing deficiency was reported. The test was conducted in a small quiet office space without any disturbance. The participants were initially asked to adjust the sound level to feel comfortable but asked to leave the level unchanged throughout the experiment session. The stimuli pairs and the corresponding references played synchronized, and the participants were allowed to listen to any of the presented stimuli as many times as they liked. The task was given as follows: “Please select the stimulus which you think sounds closer to the reference.” They were also asked to make use of the full 15 seconds of the stimuli to make their choices. Even though there were pairs with little perceptual differences, the participants were asked to try their best to select one and no tie was allowed for. A test session took 20 to 30 minutes per participant.

3 Results and analyses

The results from the listening test are described in this section. The combinations of the three magnitude separation scenarios and the two phase retrieval cases, as described in the previous section, are summarized in Table 2. Their shortened names as in the table will be used for identification hereafter.
3.1 Indirect scaling

Firstly, the pairwise comparison data from the participants was converted into estimates of similarity rating on a continuous scale, with a view to examining the rank order of the six cases compared for the test, and potentially to finding out the perceptual distances between them. The Bradley-Terry-Luce (BTL) model [24, 25] was used for this indirect scaling. A software implementation of this model in [26] was used to derive the estimates. Fig.1 shows the converted rating estimates for the six tested cases per each song.

It is seen that the FCN cases, regardless of the phase, are rated the lowest, and the REF magnitude-vocal phase combination is rated the highest in general. The ratings for the other combinations (IBM-mix, IBM-voc, and REF-mix) are not easily comparable, with some song dependency. The difference between the mix phase and the vocal phase for the FCN cases cannot be observed from the plot. This may be related to the fact that these two samples for each song sounded very similar to each other, but were clearly distinguishable from the rest due to their overall heavy distortions, which was also confirmed from the feedbacks by most of the participants after the test. This indeed weakens the applicability of the BTL model itself to the data [26], and motivates other analysis approaches, instead of attempting to make any further conclusion solely based on this indirect scaling. The next subsection details the analysis made directly upon the participants’ answers.

3.2 Win-loss tables and statistical significance

The so-called win-loss table was constructed from the participants’ answers. This indicates per pair presented in the test, how many times one of the two was selected against the other. Tables 3, 4 and 5 show the results for the songs numbered 45, 67 and 91 respectively.

In order to check whether there has been any statistically significant difference in selection between any pair, the critical number of wins, required for statis-
where $n$ is the number of subjects, $p_1$ is the proportion of the selection in population if chance alone is operating, and $z$ is the $z$-value corresponding to specified criterion of significance (type 1 error risk). When using the typical value of $\alpha = 0.05$, the corresponding $z$-value is 1.65 from the $Z$ table. Using $n = 26$ and $p_1 = 0.5$ gives $c' = 17.71$. This indicates that if a number of wins of a stimulus in a pair is equal to or larger than 18, then it can be claimed to be statistically significantly rated better than the other. The numbers below $c'$ on the above win-loss tables are highlighted. These correspond to the pairs of combinations for which the null hypothesis cannot be rejected. In the majority of cases, the distinction between the pairs was significant, with some variations depending on the song.

**Song 45**

With this song, there seems to be no confusion between any of the tested pairs. It is seen that with the same magnitude separation scenario, the vocal phase was clearly favored over the mix phase. Also, the distinction between the source separation scenarios is clear - IBM was favored over FCN, and REF over IBM, although between the REF-mix case and IBM-voc case, the REF-mix case won just above the statistically significant level.

**Song 67**

For this song, firstly it seems that the participants could not easily distinguish the FCN-voc case from the FCN-mix case. With the other magnitude separation scenarios (IBM and REF), the distinction between the mix phase and the vocal phase is clear, with the vocal phase mostly favored. Interestingly, confusions are also found between the (IBM-mix, REF-mix) pair and between the (IBM-voc, REF-mix) pair. In other words, it seems that the reference vocal magnitude with the mix phase was not judged to be better than the two IBM cases.

**Song 91**

Here a confusion between the mix phase and the vocal phase is found in the IBM scenario. In addition, the REF-mix case is not judged to be significantly better than the IBM-voc case, indicating that the use of the vocal phase with the IBM magnitude has made it hard to distinguish that from the reference magnitude with the mix phase. The (FCN-mix, FCN-voc) pair is distinguishable but again just above the significant level.
4 Discussion

The first notable finding from the results is that the difference between the mix phase and the vocal phase within the same magnitude separation scenario was perceivable in the majority of tested cases, only with the exception of FCN with song 67 and IBM with song 91. In the case of FCN particularly, as mentioned in the previous section, the distinction between the two phases had been found to be very difficult during the stimulus preparation stage. We had expected that this could be related to the overall performance of FCN being far worse than the other two cases (IBM and REF) regardless of the phase, due to the already present heavy distortions. However, having the correct phase information still seems to have improved the perceived quality for songs 45 and 91. On the contrary, in the case of IBM, during the preparation stage it had been found and expected that the difference between the mix and vocal phases would be more easily noticeable than in the FCN scenario. Nevertheless, perceptual confusion was seen between the two phases for song 91. It is thus also probable that the characteristics of the songs and of the selected 15-second excerpts had some influence on the phase difference detectability.

Another noteworthy finding is the confusion between some cases in the IBM separation scenario and the REF-mix case for songs 67 and 91. For both songs, the IBM with the vocal phase is not perceptually distinguishable by the participants from the reference magnitude-mix phase combination. This can be because 1) having the correct phase with the IBM here improved the perceived quality to be comparable to a case with perfect magnitude separation, 2) having incorrect phase degraded the perceived quality even if the magnitude separation is perfect, 3) the perceptual difference between the IBM magnitude separation and the reference magnitude was already small for these songs, or any combination of these. In addition, for song 67, the reference magnitude with the wrong phase is even confused with the IBM magnitude separation with the same wrong phase, further evidencing the disadvantage of having the incorrect phase information.

In order to reveal more comprehensive findings of the perceptual importance of the correct phase, this study could be enhanced further in these aspects:

- More songs or various segments of songs could be introduced, ideally through systematic selection, for the listening test to better observe the influence of any measurable properties of songs, such as the phase differences between the target and the mix.
- In addition to introducing the perfect phase case for comparison, other practical approaches of phase estimation can potentially be included as possible intermediate solutions.

Still, the findings from the presented experiment are meaningful in that the benefit of having the correct phase has been perceptually validated.

5 Summary

This study was conceived with the aim to investigate whether achieving the correct phase would allow for any perceivable benefit compared to the conventional audio source separation approaches in which the phase from the original mix is often used with the separated magnitude spectrum. A listening test was conducted, where the participants evaluated the stimuli corresponding to combinations of three different magnitude source separation scenarios (representing practical, close-to-ideal, and perfect separation) and two phase retrieval cases (from the original mix or from the target vocal) for three songs, by means of pairwise comparisons with external references. The results have shown that in general there is noticeable perceptual differences between the incorrect phase of the mix and the correct phase of the target vocal within the same magnitude source separation scenarios, only with a few exceptions depending on the songs. Also, it has been observed that even when perfect separation is assumed in magnitude, using the mix phase has degraded the perceived quality to be confused with that of some close-to-ideal magnitude separation cases. These findings imply that obtaining the correct phase information would provide additional perceivable benefits to the conventional magnitude-based source separation techniques.

Acknowledgement

This work was supported by Engineering and Physical Sciences Research Council (EPSRC), U.K.
Chungeun Kim et al. Perception of phase changes in the context of audio source separation

(EP/L027119/2, Musical Audio Repurposing using Source Separation). The authors also wish to thank Mr Qiuqiang Kong at CVSSP, University of Surrey for his help in conducting the listening test.

References


