PRIMARY-AMBIENT SOURCE SEPARATION FOR UPMIXING TO SURROUND SOUND SYSTEMS

Karim M. Ibrahim *

National University of Singapore
karim.ibrahim@comp.nus.edu.sg

Mahmoud Allam

Nile University
mallam@nu.edu.eg

ABSTRACT

Extracting spatial information from an audio recording is a necessary step for upmixing stereo tracks to be played on surround systems. One important spatial feature is the perceived direction of the different audio sources in the recording, which determines how to remix the different sources in the surround system. The focus of this paper is the separation of two types of audio sources: primary (direct) and ambient (surrounding) sources. Several approaches have been proposed to solve the problem, based mainly on the correlation between the two channels in the stereo recording. In this paper, we propose a new approach based on training a neural network to determine and extract the two sources from a stereo track. By performing a subjective and objective evaluation between the proposed method and common methods from the literature, the proposed approach shows improvement in the separation accuracy, while being computationally attractive for real-time applications.

Index Terms— Audio Source Separation, Primary-ambient Separation, Surround Sound Systems, Upmixing.

1. INTRODUCTION

Audio recordings are modeled as a mixture of different sources accumulated together. These sources can be divided to two different types of sources: primary (direct) and ambient (diffuse) sources. Primary sources are coherent signals that are perceived as produced from a certain direction, e.g. the main vocalist in a song. Ambient sources, e.g. reverberations, applause or crowd cheers, are uncorrelated signals perceived as sources with no certain direction, which sound surrounding. The separation of primary and ambient sources can be used in upmixing a recording, i.e. increasing the number of channels from a recording with fewer channels [1, 2]. When upmixing to systems with more channels than the original recording, extracting the ambient sources can be used to remix them into the additional channels to create the surrounding feeling supported by these sound systems. The primary sources can still be played on the same intended channels to keep the perceived directions of the different sources as originally intended in the recording[3].

Audio recordings are often mixed in a stereo two-channel mixture. The two-channel model is suitable for separating the primary and ambient sources as it resembles the human auditory model composed of two input channels, i.e. the two ears. The human brain determines the direction of the sound based on the difference between the signals reaching the left and right ears to determine the interaural time difference (ITD) and interaural level difference (ILD) [4, 5]. Hence, a stereo recording embeds enough information to simulate the human auditory system in separating the ambient sources. The key characteristic in distinguishing the ambient sources is the correlation between the signals in the two channels. Ambient sources show low correlation between the two channels, hence, the human auditory system cannot determine the direction of the source by analyzing and comparing the two signals.

The focus of this paper is to separate the primary and ambient sources from stereo mixtures, since it is the most commonly used recording technique. The paper is structured as follows: Section 2 reviews previous efforts in solving the problem and lists the limitations of these methods. Section 3 explains our proposed method to improve the separation using neural networks. Finally, Section 4 presents both objective and subjective evaluation of the proposed method with respect to the previous methods from the literature.

2. BACKGROUND

Several approaches have been proposed for the Primary-Ambient Extraction (PAE) problem in stereo recordings. A commonly used approach that has been extensively used and improved is using the Principal Component Analysis (PCA) as in the popular approach by Goodwin [6, 7]. PCA is a suitable approach for the problem as it uses the correlation between the two channels to extract the correlated signals from the mixture as the primary source while ambient sources are assumed to be the residuals, which show low correlation. PCA is suitable for extracting primary sources with intensity difference between the two channels, however, it fails to make use of the time difference information. In [8], a PCA-based approach was proposed that additionally analyzes the time shift between the two channels in the separation. Another drawback in the PCA-based approaches is its low accuracy in separating the primary/ambient sources when there is no prominent primary source, i.e. when the recording is mainly ambient sources. Recently, a new PCA-based approach was proposed in [9] to improve the accuracy of separating ambient sources by using weighting factor to estimate the presence of a dominant primary source.

Another approach for the problem was proposed by Faller [10] based on using the least square method to estimate the primary and ambient sources to minimize the errors between the extracted signals and the original stereo input. A spectral-based approach was proposed by Avendano [11] to calculate a band-wise inter-channel short-time coherence. Using the cross- and autocorrelation between the stereo channels, he calculated the basis for the estimation of a panning and ambience index. A method based on separating the ambient sources using an adaptive filter algorithm to detect correlated and uncorrelated signals is proposed in [12].

Though most of the approaches are proposed for stereo recordings, there are approaches aimed at separating the sources in mono recordings. A method based on non-negative matrix factorization

*The author completed most of this work while at Nile University

(NMF) is described in [13]. Another approach for mono recordings which is based on supervised learning and low-level features extraction is presented in [14]. This approach is similar to our proposed method in using trained neural networks, however, it is intended for mono recordings and used a different set of features that suits mono recordings. It is rather limited to extracting ambient reverberations only due to the limiting nature of mono recordings.

3. NEURAL NETWORK APPROACH

In this paper, we consider the primary/ambient extraction task as a classification problem, where we classify each frequency-frame bin to be either primary or ambient, and then to reconstruct the two signals based on the classification using a trained neural network. In this section, we examine the process of setting-up and using the neural network for the intended task.

3.1. The Setup

The three main steps for setting-up this neural network are: collecting a reliable dataset of primary/ambient sources, training the neural network and applying the classification on the target recordings.

3.1.1. The Dataset

In order to ensure having a reliable separation, we need to ensure that the data we use for training the neural network is reliable and well-labeled. The separation will be highly dependent on the data we use for training. For the primary-ambient separation, we need to use data that represents the sources precisely and spans over a large variety of sound sources to ensure that the neural network learns to discriminate between sources from different setups.

We collected the dataset using recordings from Apple Loops. We particularly selected recordings tagged with dry, i.e. no reverberations or effects added, to be primary sources. We selected recordings labeled with reverberations and sound effects to be ambient. We then went through an additional phase of filtering by listening to the selected excerpts and ensure they sound either completely primary or ambient. We selected 280 excerpts divided equally between primary and ambient sources, with a length of 15 seconds for every excerpt. Samples for primary sources included solo music instruments, human voices in a dialogue and animal sounds. Samples of ambient sources included sounds effects as forests, rain, traffic, cheering crowd, echo and reverberations. All the sources are labeled as either primary sources that do not include any reverberations or surrounding effects or, conversely, labeled as ambient. The dataset is divided to 200 excerpts for training and 80 excerpts for testing.

The next step is to extract the feature vectors from the dataset using the following steps:

1. Starting from the original two-channel signals, $x_l[n]$ and $x_r[n]$, we apply the STFT on the signals to get $X_l[m,k]$ and $X_r[m,k]$. We calculate the STFT using $\frac{3}{4}$ overlapping Hamming windows of 4096 samples, corresponding to a duration of 92.8 milliseconds at a sampling frequency of 44.1 kHz.
2. We clean the data by removing the frames in the STFT domain that contain an energy level less than the average energy level of the input file by 30 dB. This is to remove the frames that have negligible information, as they do not have a large impact on the training process.
3. The feature vectors are the STFT values of each frequency-frame bin combined with two preceding and two succeeding bins for both channels to get temporal context, experiments showed that two frames gives as good results as taking additional frames.
4. Since the STFT values are complex, we split them into real and imaginary values, ending up with a single feature vector as shown in Figure 1.

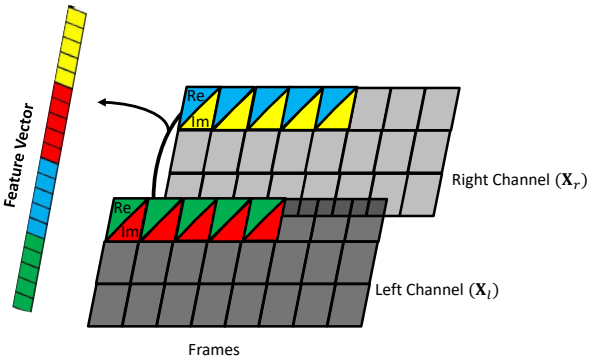

Fig. 1. Extracting the feature vectors of the STFT of the input signal.

3.1.2. Training the network

The next step is to train a fully connected feed-forward neural network using the data we collected to fit the PAE model. The training parameters of the network were chosen empirically. The network is made of 3 hidden layers of size 15,10 and 2 nodes respectively. All layers use a rectified linear unit (ReLu) as an activation function. The last layer's output range between 1 and 0 using Sigmoid activation function to represent the probability of the source being primary. We trained the network using batch gradient decent running for 200 epochs and using sum square error as cost function.

3.1.3. Applying the separation

The final step is to apply the neural network on the target input to be separated to the primary and ambient components. We use the neural network to predict the probability of each frequency-frame of the input file to be primary, then we form a mask of values between 0 and 1 in the time-frequency domain that corresponds to the prediction. Finally, by multiplying the mask to the input STFT we extract the primary component in the time-frequency domain, similarly by applying the complement of the mask we extract the ambient component.

4. EVALUATION

In this section, we discuss the evaluation of different primary-ambient separation methods. We perform two evaluation methods to measure the accuracy of the extraction, one is subjective, based on the user experience. The second is objective, based on the performance measurements used for blind source separation described in [15] and adapted for the problem of primary/ambient extraction in [9].

4.1. Subjective Evaluation

The first part of the evaluation is based on the user experience. We performed two experiments; the first is to evaluate the different playback systems to determine the utility of PAE, and the second is to evaluate the different PAE methods. Both of the experiments were done under the following conditions:

1. The systems were played in a random order
2. The participants did not know what system was being played nor did they know what the different systems were.
3. The participants were asked to order the systems in terms of the most surrounding and appealing sound.
4. Total number of participants: 11
5. The playback setup was made up from 4 surround speakers equally spaced from the participant. as shown in Figure 2.
6. For each system two songs (each of length 30 seconds) were played. We selected songs that contain high ambience and induce a surrounding feeling so that would enable the participants to evaluate the surround sound systems. The songs are:
   (a) Diamonds on the Soles of Her Shoes by Paul Simon.
   (b) Rock You Gently by Jennifer Warnes.
7. All the systems were adjusted to have the same energy level at the spot where the participant is sitting.

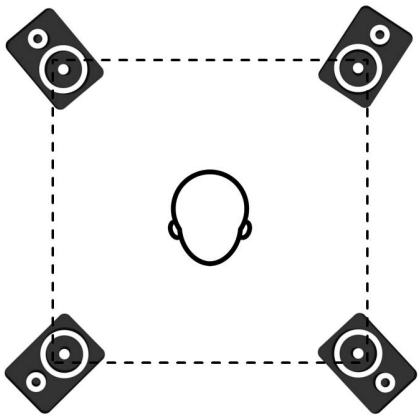

Fig. 2. Experiment's playback system arrangement

4.1.1. Experiment 1: Different Playback Systems:

The point of this experiment is to evaluate the different arrangements of sound systems and to test whether users sense and appreciate surround systems compared to traditional sound systems, which in turn justifies the need for using primary-ambient separation for upmixing. The different systems are: mono single-channel system rendered by duplicating the input on both front channels, referred to as `Mono`, stereo two-channel system, referred to as `Stereo`, 4-channel system, stereo played on front speakers and same stereo played on back speakers, referred to as `4CH Stereo`, 4-channels system, primary played on front speakers and ambient played on back speakers, referred to as `Ambient Back` and 4-channels system, primary played on front speakers and ambient played on all speakers, referred to as `Ambient all`.

Mono	Stereo	4CH Stereo	Ambient Back	Ambient All
5	3	2	4	1
2	5	4	1	3
4	3	1	5	2
5	4	3	1	2
5	3	1	4	2
5	4	3	2	1
5	3	4	2	1
5	4	3	1	2
4	5	2	3	1
4	3	5	2	1
5	4	1	3	2
4.5	3.7	2.6	2.5	1.6

Table 1. Rating of the different playback systems

Table 1 shows the ratings of the 11 participants (where 1 is the most favorite and 5 is the least favorite), while the last row represents the average of the ratings. The selected participants had experience in critical listening and were familiar the concepts of spatial sound. We find that most participants picked the Mono system to be their least favorite as expected, this was acting as an anchor for the experiment to make sure the results are sensible. We find that the stereo and the 4-channels stereo are judged as the least favorite after mono. The primary-ambient separation was picked to be the most preferred system, which concludes that the separation makes an improvement in the playback systems. The system where the ambient is being played on all speakers is favored over the one where the ambient is played only in the back, this was expected since the ambient sources should be perceived as coming from all around.

4.1.2. Experiment 2: Different Separation Methods:

This experiment was made to evaluate the different separation methods based on the user-experience and to test if the objective evaluation agrees with the actual users' preference. The different PAE methods selected are popular methods from literature that were accessible during the experiment. The methods are: The neural network method proposed in this paper, The modified PCA method by Goodwin in [6, 7], The extraction method by Avendano in [11] and The panning-estimation-based method by Kraft and Zlzer in [16].

Table 2 shows the rating of 10 participants, one participant could not feel any difference between the methods. Similar to the previous experiment, 1 is picked for the most favorite method. We find that, according to the users' preference, the neural network method is the most favorite in terms of being surrounding and appealing, followed by the PCA-based method by Goodwin. This shows that, perceptually, the neural network separation is more preferred by users than the previously proposed methods.

4.2. Objective Evaluation

The objective evaluation is based on the "BSS Eval" toolbox proposed in [15] which is intended to evaluate blind audio source separation (BASS). However, an adaptation for the primary/ambient separation was proposed in [9], which is used in this paper to evaluate the neural network with different methods from the literature.

Neural Network	PCA by Goodwin	Avendano	Panning Estimation
3	2	1	4
4	2	1	3
1	3	4	2
1	4	3	2
1	2	4	3
1	2	4	3
1	2	3	4
2	3	1	4
1	2	4	3
1	3	4	2
1.6	2.5	2.9	3.0

Table 2. Rating of the different PAE methods

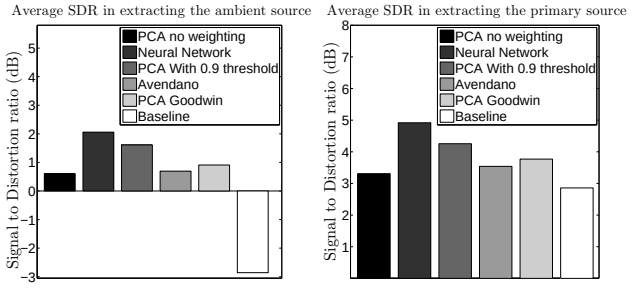

Fig. 3. Average SDR in primary and ambient extraction

As explained in [9], the "BSS Eval" method can be adapted to the problem of PAE by composing a mixture of two sources, one is all ambient and one is all primary. In the ideal case, applying a PAE method would separate two sources identical to the originals. However, due to the limitations of the extraction methods, there is interference between the two sources. Hence, this error can be measured using the metrics in the "BSS Eval" toolbox.

The evaluation is performed on five different PAE methods: The Principal Component Analysis (PCA) without adding weighting, referred to as `PCA without weighting`, the neural network method proposed in this paper, referred to as `Neural Network`, PCA-based approach with adaptive weighting proposed in [9], using 0.9 threshold, referred to as `PCA Adaptive`, the extraction method by Avendano and Jot in [11], referred to as `Avendano` and the weighted PCA method by Goodwin in [6, 7]. Referred to as `PCA Goodwin`. Audio samples of the different methods are available online[1].

The evaluation was performed using two datasets, one is made out of all ambient sources and the second is made of all primary sources. The total number of mixed sources is 40 of each type. We used the Matlab toolbox "BSS Eval" [17] for calculating the errors. The evaluation was made out as follows:

1. Mixing one ambient source with one primary source after normalizing the two of them.
2. Applying the five different PAE methods to extract the primary and ambient sources.
3. Use the extracted outputs and the original sources to evaluate each method.
4. A baseline is defined by comparing the original ambient or primary sources to the mixture without any separation. This is used to define the improvement of each extraction method over the original mixture.

Figure 3 shows the average Signal to Distortion ratio (SDR) in extracting both the primary and the ambient sources for different methods. By analyzing the graph, we find that the neural network improves the separation quality for both the primary and ambient sources over both popular methods as `Avendano` and `PCA Goodwin` and recent methods as the `PCA Adaptive`. The objective evaluation results matches the preferences of the users obtained from the subjective evaluation. This emphasizes the validity of the objective evaluation method proposed in [9] and used in the paper.

[1] http://www.comp.nus.edu.sg/%7ekarim/PAE/PAE.html

5. CONCLUSIONS

According to both the subjective and objective evaluation, we find that the neural network performs significantly better than the previously suggested methods. This is perceived in terms of the accuracy of separating the primary and ambient sources and producing an appealing surround sound. The subjective evaluation also showed that using the PAE separation improves the sound system and is preferred by the users over the original typical playback systems.

6. REFERENCES

[1] Ville Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*. Audio Engineering Society, 2006.

[2] Mingsian R Bai and Geng-Yu Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *Consumer Electronics, IEEE Transactions on*, vol. 53, no. 3, pp. 1011–1019, 2007.

[3] Derry Fitzgerald, "Upmixing from mono-a source separation approach," in *Digital Signal Processing (DSP), 2011 17th International Conference on*. IEEE, 2011, pp. 1–7.

[4] Arthur N Popper and Richard R Fay, *Sound source localization*, Springer, 2005.

[5] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[6] Michael M Goodwin and J-M Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 1, pp. I–9.

[7] Michael M Goodwin, "Geometric signal decompositions for spatial audio enhancement," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 409–412.

[8] Jianjun He, Ee-Leng Tan, and Woon-Seng Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 266–270.

[9] Karim M. Ibrahim and Mahmoud Allam, "Primary-ambient extraction in audio signals using adaptive weighting and principal component analysis," in *Proceedings of the 13th Sound and*

Music Computing Conference (SMC), Hamburg, Germany, 2016, pp. 227–232.

[10] Christof Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 54, no. 11, pp. 1051–1064, 2006.

[11] Carlos Avendano and Jean-Marc Jot, "A frequency-domain approach to multichannel upmix," *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 740–749, 2004.

[12] John Usher, Jacob Benesty, et al., "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *Ieee Transactions on Audio Speech and Language Processing*, vol. 15, no. 7, pp. 2141, 2007.

[13] Christian Uhle, Andreas Walther, Oliver Hellmuth, and Juergen Herre, "Ambience separation from mono recordings using non-negative matrix factorization," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[14] Christian Uhle and Christian Paul, "A supervised learning approach to ambience extraction from mono recordings for blind upmixing," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx08), Espoo, Finland*, 2008, pp. 137–144.

[15] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] Sebastian Kraft and Udo Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *18th International Conference on Digital Audio Effects (DAFx)*, 2015.

[17] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.