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.

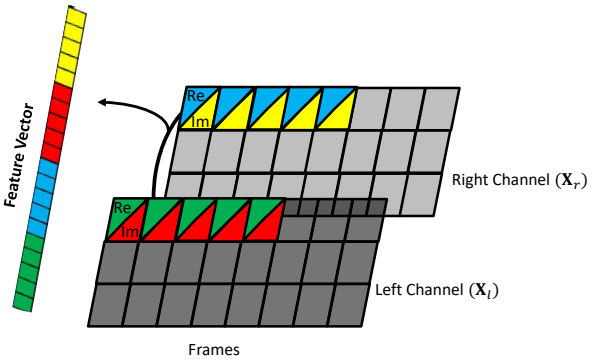

Fig. 1. Extracting the feature vectors of the STFT of the input signal.

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.

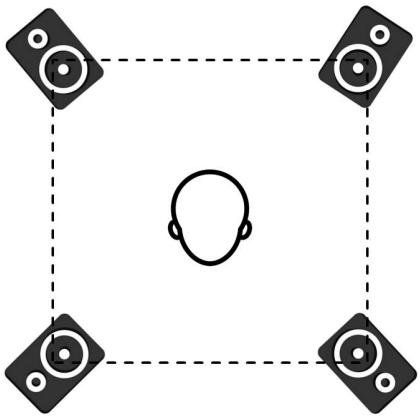

Fig. 2. Experiment's playback system arrangement

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

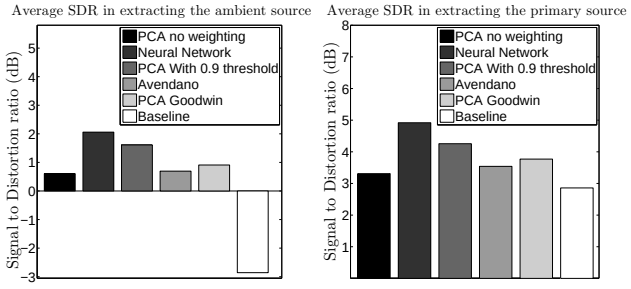

Fig. 3. Average SDR in primary and ambient extraction

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.

INTELLIGIBILITY OF SUNG LYRICS: A PILOT STUDY

Karim M. Ibrahim[1] **David Grunberg**[1] **Kat Agres**[2]
Chitralekha Gupta[1] **Ye Wang**[1]
[1] Department of Computer Science, National University of Singapore, Singapore
[2] Institute of High Performance Computing, A*STAR, Singapore
karim.ibrahim@comp.nus.edu.sg, wangye@comp.nus.edu.sg

ABSTRACT

We propose a system to automatically assess the intelligibility of sung lyrics. We are particularly interested in being able to identify songs which are intelligible to second language learners, as such individuals often sing along the song to help them learn their second language, but this is only helpful if the song is intelligible enough for them to understand. As no automatic system for identifying 'intelligible' songs currently exists, songs for second language learners are generally selected by hand, a time-consuming and onerous process. We conducted an experiment in which test subjects, all of whom are learning English as a second language, were presented with 100 excerpts of songs drawn from five different genres. The test subjects listened to and transcribed the excerpts and the intelligibility of each excerpt was assessed based on average transcription accuracy across subjects. Excerpts that were more accurately transcribed on average were considered to be more intelligible than those less accurately transcribed on average. We then tested standard acoustic features to determine which were most strongly correlated with intelligibility. Our final system classifies the intelligibility of the excerpts and achieves 66% accuracy for 3 classes of intelligibility.

1. INTRODUCTION

While various studies have been conducted on singing voice analysis, one aspect which has not been well-studied is the *intelligibility* of a given set of lyrics. Intelligibility describes how easily a listener can comprehend the words that a performer sings; the lyrics of very intelligible songs can easily be understood, while the lyrics of less intelligible songs sound garbled or even incomprehensible to the average listener. People's impressions of many songs are strongly influenced by how intelligible the lyrics are, with one study even finding that certain songs were perceived as 'happy' when people could not understand its lyrics, but was perceived as 'sad' when the downbeat lyrics were made comprehensible [20]. It would thus be useful to enable systems to automatically determine intelligibility, as it is a key factor in people's perception of a wide variety of songs.

We are particularly interested in measuring the intelligibility of songs with respect to second language learners. Many aspects of learning a second language to the point of fluency have been shown to be difficult, including separating the phonemes of an unfamiliar language [30], memorizing a large number of vocabulary words and grammar rules [22], and maintaining motivation for the length of time required to learn the language. Consequently, many second language learners need help, and music has been shown to be a useful tool for this purpose. Singing and language development have been shown to be closely related at the neurological level [24, 32], and experimental results have demonstrated that singing along with music in the second language is an effective way of improving memorization and pronunciation [12, 19]. However, specific songs are only likely to help these students if they can understand the content of the lyrics [11]. As second language learners may have difficulty understanding certain songs in their second language due to their lack of fluency, they could be helped by a system capable of automatically determining which songs they are likely to find intelligible or unintelligible.

We therefore seek to design a system which is capable of assessing a given song and assigning it an intelligibility score, with the standard of intelligibility biased towards people who are learning the language of the lyrics but have not yet mastered it. To gather data for this system we compiled excerpts from 50 songs and had volunteering participants listen to the song in order to discover how intelligible they found the lyrics. Rather than simply having the participants rate the intelligibility of the song, we had the participants transcribe the lyrics that they heard and then calculated an intelligibility score for each excerpt based on the statistics of how accurately the students transcribed it. Excerpts that were transcribed more accurately on average were judged to be more intelligible than those transcribed less accurately on average. A variety of acoustic features were then used to build a classifier which could determine the intelligibility of a given piece of music. The classifier was then run on the same excerpts used in the listening experiment, and the results of each were compared.

The remaining outline of this paper is as follows: Sec-

tion 2 lists relevant literature in the field. Section 3 describes the transcription experiment performed to gather data. Section 4 discusses the features and the classifier. Finally, Sections 5 and 6 shows the evaluation of our proposed model and our conclusions, respectively.

2. LITERATURE REVIEW

That sung lyrics could be more difficult to comprehend than spoken words has long been established in the scientific community. One study showed that even professional voice teachers and phoneticians had difficulty telling vowels apart when sung at high pitch [7]. Seminal work by Collister and Huron found listeners to make hearing errors as much as seven times more frequently when listening to sung lyrics than spoken ones [3]. Such studies also noted lyric features which could help differentiate intelligible from unintelligible songs; for instance, one study noted that songs comprised mostly of common words sounded more intelligible than songs with rarer words [9]. However, lyric features alone are not sufficient to assess intelligibility; the same lyrics can be rendered more or less intelligible depending on, for instance, the speed at which they are sung. These other factors must be taken into account to truly assess lyric intelligibility.

Studies have been conducted on assessing the overall *quality* of singing voice. One acoustic feature which multiple studies have found to be useful for this purpose is the power ratio of frequency bands containing energy from the singing voice to other frequency bands; algorithms using this feature have been shown to reliably distinguish between trained and untrained singers [2,23,34]. Calculation of pitch intervals and vibrato have also been shown to be useful for this purpose [21]. However, while the quality of singing voice may be a factor in assessing intelligibility, it is not the only such factor. Aspects of the song that have nothing to do with the skill of the singer or the quality of their performance, such as the presence of loud background instruments, can contribute, and additional features that take these factors into account are needed for a system which determines lyric intelligibility.

Another related task is that of singing transcription, in which a computer must listen to and transcribe sung lyrics [18]. It may seem that one could assess intelligibility by comparing a computer's transcription of the lyrics to a ground truth set of lyrics and determining if the transcription is accurate. But this too does not really determine intelligibility, at least as humans perceive it. A computer can use various filters and other signal processing or machine learning tools to process the audio and make it easier to understand, but a human listening to the music will not necessarily have access to such tools. Thus, even if a computer can understand or accurately transcribe the lyrics of a piece of music, this does not indicate whether those lyrics would be intelligible to a human as well.

3. BEHAVIORAL EXPERIMENT

To build a system that can automatically process a song and evaluate the intelligibility of its lyrics, it is essential to gather ground truth data that reflects this intelligibility on average across different listeners. Hence, we conducted a study where participants were tasked with listening to short excerpts of music and transcribing the lyrics, a common task for evaluating intelligibility of lyrics [4]. The accuracy of their transcription can be used to assess the intelligibility of each excerpt.

3.1 Method

3.1.1 Participants

Seventeen participants (seven females and ten males) volunteered to take part in the experiment. Participants were between 21 to 41 years (mean = 27.4 years). All participants indicated no history of hearing impairment and that they spoke some English as a second language. Participants were rewarded with a $10 voucher for their time. Participants were recruited through university channels via posters and fliers. The majority of the participants were university students.

3.1.2 Materials

For the purpose of this study, we focused solely on English-language songs. Because one of the main applications for such a system is to recommend music for students who are learning foreign languages, we focused on genres that are popular for students. To identify these genres, we asked 48 university students to choose the 3 genres that they listen to the most, out of the 12 genres introduced in [4], as these 12 genres cover a wide variety of singing styles. The twelve genres are: Avante-garde, Blues, Classical, Country, Folk, Jazz, Pop/Rock, Rhythm and Blues, Rap, Reggae, Religious, and Theater. Because the transcription task is long and tiring for participants, we limited the number of genres tested to only five, from which we would draw approximately 45 minutes worth of music for transcription. We selected the five most popular genres indicated by the 48 participants: Classical, Folk, Jazz, Pop/Rock, and Rhythm and Blues.

After selecting the genres, we collected a dataset of 10 songs per genre. Because we were interested in evaluating participants' ability to transcribe an unfamiliar song, as opposed to transcribing a known song from memory, we focused on selecting songs that are not well-known in each genre. We approached this by selecting songs that have less than 200 ratings on the website Rate Your Music (rateyourmusic.com). Rate Your Music is a database of popular music where users can rate and review different songs, albums and artists. Popular songs have thousands of ratings while less known songs have few ratings. We used this criteria to collect songs spanning the 5 genres to produce our dataset. The songs were randomly selected, with no control over the vocal range or the singer's accent, as long as they satisfied the condition of being in English and having few ratings.

Because transcribing an entire song, let alone 50 songs, would be an overwhelming process for the participants, we selected short excerpts from each song to be transcribed. Two excerpts per song were selected randomly such that each excerpt would include a complete utterance (e.g., no excerpts were terminated mid-phrase). Excerpts varied between 3 to 16 seconds in length (average = 6.5 seconds), and contained 9.5 words on average. The ground-truth lyrics for these songs were collected from online sources and reviewed by the experimenters to ensure they matched the version of the song used in the experiment. It is important to note that selecting short excerpts might affect intelligibility, because the context of the song (which may help in understanding the lyrics) is lost. However, using these short excerpts is essential in making the experiment feasible for the participants, and would still broadly reflect the intelligibility of the song. The complete dataset is composed of 100 excerpts from 50 songs, 2 excerpts per song, covering 5 genres, and 10 songs per genre. Readers who are interested in experimenting on the dataset can contact the authors.

3.1.3 Procedure

We conducted the experiment in three group listening sessions. During each session, the participants were seated in a computer lab, and recorded their transcriptions of the played excerpts on the computer in front of them. The excerpts were played in randomized order, and each excerpt was played twice consecutively. Between the two playbacks of each excerpt there was a pause of 5 seconds, and between different excerpts a pause of 10 seconds, to allow the participants sufficient time to write their transcription. The total duration of the listening session is 46:59 minutes. Two practice trials were presented before the experimental trials began, to familiarize participants with the experimental procedure.

3.2 Results and Discussion

To evaluate the accuracy of the participants' transcription, we counted the number of words correctly transcribed by the participant that match the ground truth lyrics. For each transcription by each student, the ratio between correctly transcribed words to the total number of words in the excerpt was calculated. We then calculated the average ratio for each excerpt across all 17 participants to yield an overall score for each excerpt between 0 and 1. This score was used to represent the ground-truth transcription accuracy, or *Intelligibility score*, for each excerpt. The distribution of Intelligibility scores in the dataset is shown in Figure 1. From the figure, we can observe that the intelligibility scores are biased towards higher values, i.e. there are relatively few excerpts with a low intelligibility score. This may be caused by the restricted set of popular genres indicated by students, as certain excluded genres would be expected to have low intelligibility, such as Heavy Metal.

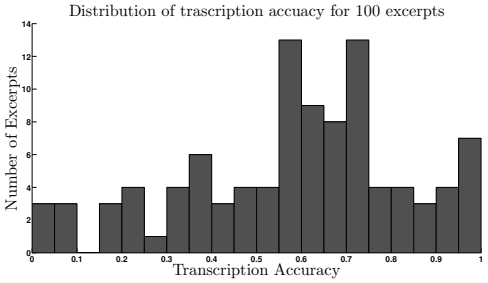

Figure 1. The distribution of the transcription accuracies (Intelligibility score).

4. COMPUTATIONAL SYSTEM

The purpose of this study is to select audio features that can be used to build a system capable of 1) predicting the intelligibility of song lyrics, and 2) evaluating the accuracy of these predictions with respect to the ground truth gathered from human participants. In the following approach, we analyze the input signal and extract expressive features that reflect the different aspects of an intelligible singing voice. Several properties may contribute to making the singing voice less intelligible than normal speech. One such aspect is the presence of background music, as accompanying music can cover or obscure the voice. Therefore, highly intelligible songs would be expected to have a dominant singing voice compared with the accompanying music [4]. Unlike speech, the singing voice has a wider and more dynamic pitch range, often featuring higher pitches in soprano vocal range. This has been shown to affect the intelligibility of the songs, especially with respect to the perception of sung vowels [1, 3]. An additional consideration is that in certain genres, such as Rap, singing is faster and has a higher rate of words per minute than speech, which can reduce intelligibility. Furthermore, as indicated in [10], the presence of common, frequently occurring words helps increase intelligibility, while uncommon words decrease the likelihood of understanding the lyrics. In our model, we aimed to include features that express these different aspects to determine the intelligibility of song lyrics across different genres. These features are then used to train the model to accurately predict the intelligibility of lyrics in the dataset, based on the ground truth collected in our behavioral experiment.

4.1 Preprocessing

To extract the proposed features from an input song, two initial steps are required: separating the singing voice from the accompaniment, and detecting the segments with vocals. To address these steps, we selected the following approaches based on current state-of-the-art methods:

4.1.1 Vocals Separation

Separating vocals from accompaniment music is a well-known problem that has received considerable attention in the research community. Our approach makes use of the popular Adaptive REPET algorithm [16]. This algorithm is

based on detecting the repeating patten in the song, which is meant to represent the background music. Separating the detected pattern leaves the non-repeating part of the song, meant to capture the vocals. Adaptive REPET also has the advantage of discovering local repeating patterns in the song over the original REPET algorithm [26]. Choosing Adaptive REPET was based on two main advantages: The algorithm is computationally attractive, and it shows competitive results compared to other separation algorithms, as shown in the evaluation of [14].

4.1.2 Detecting Vocal Segments

Detecting vocal and non-vocal segments in the song is an important step in extracting additional information about the intelligibility of the lyrics. Various approaches have been proposed to perform accurate vocal segmentation, however, it remains a challenging problem. For our approach, we implemented a method based on extracting the features proposed in [15], then training a Random Forest classifier using the Jamendo corpus [1] [27]. The classifier was then used to binary classify each frame of the input file as either vocals or non-vocals.

4.2 Audio features

In this section, we investigate the set of features we used in training the model for estimating lyrics intelligibility. We use a mix of features reflecting specific aspects of intelligibility plus common standard acoustic features. The selected features are:

1. **Vocals to Accompaniment Music Ratio (VAR)**: Defined as the energy of the separated vocals divided by the energy of the accompaniment music. This ratio is computed only in segments where vocals are present. This feature reflects how strong the vocals are compared to the accompaniment. High VAR suggests that vocals are relatively loud and less likely to be obscured by the music. Hence, higher VAR counts for higher intelligibility. This feature is particularly useful in identifying songs that are unintelligible due to loud background music which obscures the vocals.

2. **Harmonics-to-residual Ratio (HRR)**: Defined as the the energy in a detected fundamental frequency (f0) according to the YIN algorithm [5] plus the energy in its 20 first harmonics (a number chosen based on empirical trials), all divided by the energy of the residual. This ratio is also applied only to segments where vocals are present. Since harmonics of the detected f0 in vocal segments are expected to be produced by the singing voice, this ratio, like VAR, helps to determine whether the vocals in a given piece of music are stronger or weaker than the background music which might obscure it.

3. **High Frequency Energy (HFE)**: Defined as the sum of the spectral magnitude above 4kHz,

$$HFE_n = \sum_{k=f_{4k}}^{N_b/2} a_{n,k} \tag{1}$$

where $a_{n,k}$ is the magnitude of block n and FFT index k of the short time Fourier transform of the input signal, f_{4k} is the index corresponding to 4 kHz and N_b is the FFT size [8]. We calculate the mean across all frames of the separated and segmented vocals signal, as we are interested in the high energy component in vocals and not the accompanying instruments. We get a scalar value per input file reflecting high frequency energy. Singing in higher frequencies has been proven to be less intelligible than music in low frequencies [3], so detection of high frequency energy can be a useful clue that such vocals might be present and could reduce the intelligibility of the music, such as frequently happens with opera music.

4. **High Frequency Component (HFC)**: Defined as the sum of the amplitudes and weighted by the frequency squared,

$$HFC_n = \sum_{k=1}^{N_b/2} k^2 a_{n,k} \tag{2}$$

where $a_{n,k}$ is the magnitude of block n and FFT index k of the short time Fourier transform of the input signal and N_b is the FFT size [17]. This is another measure of high frequency content.

5. **Syllable Rate**: Singing at a fast pace while pronouncing several syllables over a short period of time can negatively affect the intelligibility [6]. In the past, Rao et al. used temporal dynamics of timbral features to separate singing voice from background music [28]. These features showed more variance over time for singing voice, while being relatively invariant to background instruments. We expect that these features will also be sensitive to the syllable rate in singing. We use the temporal standard deviation of two of their timbral features: sub-band energy (SE) in the range of ([300-900 Hz]), and sub-band spectral centroid (SSC) in the range of ([1.2-4.5 kHz]), defined as

$$SSC = \frac{\sum_{k=k_{low}}^{k_{high}} f(k)|X(k)|}{\sum_{k=k_{low}}^{k_{high}} |X(k)|} \tag{3}$$

$$SE = \sum_{k=k_{low}}^{k_{high}} |X(k)|^2 \tag{4}$$

where $f(k)$ and $|X(k)|$ are frequency and magnitude spectral value of the k^{th} frequency bin, and k_{low} and k_{high} are the nearest frequency bins to the lower and upper frequency limits on the sub-band respectively.

[1] http://www.mathieuramona.com/wp/data/jamendo/

According to [28], SE enhances the fluctuations between voiced and unvoiced utterances, while SSC enhances the variations in the 2^{nd}, 3^{rd} and 4^{th} formants across phone transitions in the singing voice. Hence, it is reasonable to expect high temporal variance of these features for songs with high syllable rate, and vice versa. Thus, this feature is able to differentiate songs with high and low syllable rates. We would expect that very high and very low syllable rates should lead to low intelligibility score, while rates in a similar range to that of speech should result in high intelligibility score.

6. **Word-Frequency Score**: Songs which use common words have been shown to be more intelligible than those which use unusual or obscure words [10]. Hence, we calculate a word-frequency score for the lyrics of the songs as an additional feature. This feature is a non-acoustic feature that is useful in cases where the lyrics of the song are available. We calculate the word-frequency score using the `wordfreq` open-source toolbox [31] which provides an estimates of the frequencies of words in many languages.

7. **Tempo and Event Density**: These two rhythmic features reflect how fast the beat and rhythm of the song are. Event density is defined as the average frequency of events, i.e., the number of note onsets per second. Songs with very fast beats and high event density are likely to be less intelligible than slower songs, since the listener has less time to process each event before the next one begins. We used the `MIRToolbox` [13] to extract these rhythmic features.

8. **Mel-frequency cepstral coefficients (MFCCs)**: MFCCs approximates the human auditory system's response more closely than the linearly-spaced frequency bands [25]. MFCCs have been proven to be effective features in problems related to singing voice analysis [29], and so were considered as a potential feature here as well. For our system, we selected the 17 first coefficients (excluding the 0th) as well as the deltas of those features, which proved empirically to be the best number of coefficients. The MFCCs are extracted from the original signal without separation, as it reflects how the whole song is perceived.

By extracting this set of features for an input file, we end up with a vector of 43 features to be used in estimating the intelligibility of the lyrics in this song.

4.3 Model training

We used the dataset and ground-truth collected in our behavioral experiment to train a Support Vector Machine model to estimate the intelligibility of the lyrics. To categorize the intelligibility to different levels that would match a language student's fluency level, we divided our dataset to three classes:
`High Intelligibility`: excerpts with transcription accuracy of greater than 0.66.
`Moderate Intelligibility`: excerpts with transcription accuracy between 0.33 and 0.66 inclusive.
`Low Intelligibility`: excerpts with transcription accuracy of less than 0.33.
Out of the 100 samples in our dataset, 43 are in the High Intelligibility class, 42 are in the Moderate Intelligibility class, and the remaining 15 are in the Low Intelligibility class. For this pilot study, we tried a number of common classifiers, including Support Vector Machine (SVM), random forest and k-nearest neighbors. Our trials for finding a suitable model led to using SVM with a linear kernel, as it is an efficient, fast and simple model which is suitable for this problem. Finally, as a preprocessing step, we normalize all the input feature vectors before passing them to the model to be trained.

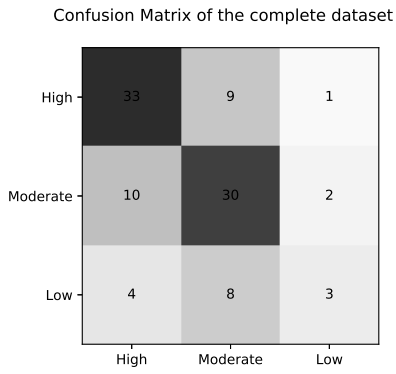

Figure 2. Confusion Matrix of the SVM output.

5. MODEL EVALUATION

Because this problem has not been addressed before in the literature, and it is not possible to perform evaluation using other methods, we based our evaluation on classification accuracy from the dataset. Given the relatively small number of samples in the dataset, we used leave-one-out cross-validation for evaluation. To evaluate the performance of our model, we compute overall accuracy, as well as the Area Under the ROC Curve (AUC). We scored AUC of 0.71 and accuracy of 66% with the aforementioned set of features and model. The confusion matrix of validating our model using leave-one-out cross-validation on our collected dataset is shown in Figure 2. The figure shows that the classifier has relatively more accuracy in predicting high and moderate than low intelligibility, which is often confused with the moderate class. Given that our findings are based on a relatively small segment of excerpts with low intelligibility, the classifier was found to be trained to work better on the high and moderate excerpts.

Following model evaluation on the complete dataset, we were interested in investigating how the model performs on different genres, specifically how it performs when tested

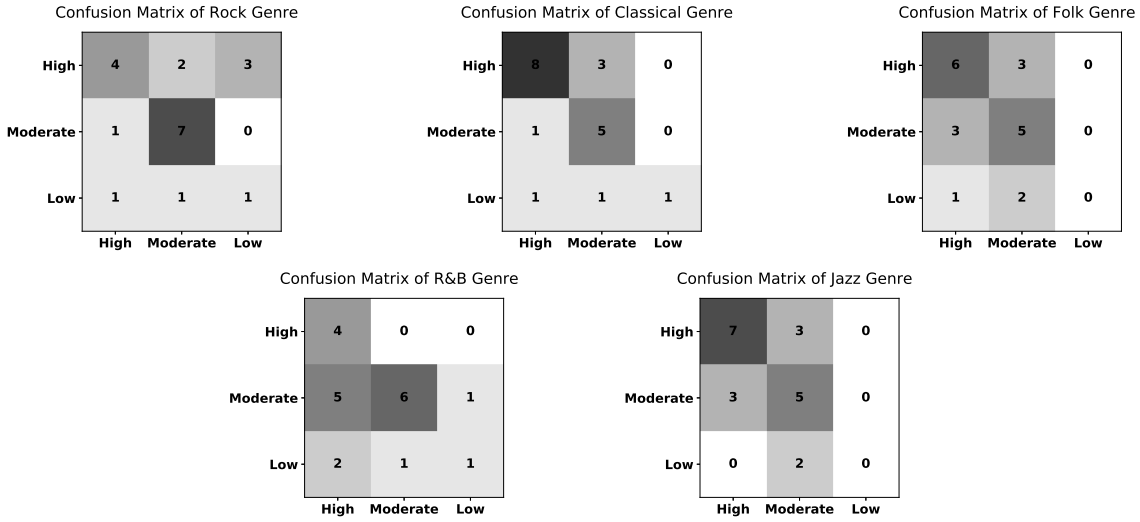

Figure 3. Confusion matrix of the different genres

Genre	Classification Accuracy
Pop/Rock	60%
R&B	55%
Classical	70%
Folk	55%
Jazz	60%

Table 1. Classification accuracy for different genres

with a genre that was not included in the training dataset. This would imply how the model generalizes when running on different genres that was not present during training, as well as showing how changing genres affect classification accuracy. We performed an evaluation where we trained our model using 4 out of the 5 genres in our dataset, and tested it on the 5th genre. The classification accuracy across different genres is shown in Table 1. The results show variance in classifying different genres. For example, Classical music receives higher accuracy, while genres as Rhythm and Blues and Folk shows less accuracy. By analyzing the confusion matrices of each genre shown in Figure 3, we found that the confusion is mainly between high and moderate classes.

By reviewing the impact of the different features on the classifier performance, we looked into what features have the biggest impact using the attribute ranking feature in Weka [35]. We found that several MFCCs contribute most in differentiating between the three classes, which we interpret to be due to analyzing the signal in different frequency sub-bands incorporates perceptual information of both the singing voice and the background music. This was followed by the features reflecting the syllable rate in the song, because singing rate can radically affect the intelligibility. Vocals-to-Accompaniment Ratio and High Frequency Energy followed in their impact on differentiating between the three classes. The features that had the least impact were the tempo and event density, which does not necessarily reflect the rate of singing.

For further studies on the suitability of the features in classifying songs with very low intelligibility, the genres pool can be extended to include other genres with lower intelligibility, rather than being limited to the popular genres between students. Further studies can also include the feature selection and evaluation process: similar to the work in [33], deep learning methods may be explored to select the features which perform best, rather than hand-picking features, to find the most suitable set of features for this problem. It is possible to extend the categorical approach of intelligibility levels to a regression problem, in which the system evaluates the song's intelligibility with a percentage. Similarly, certain ranges of the intelligibility score can be used to recommend songs to students based on their fluency level.

6. CONCLUSION

In this study, we investigated the problem of evaluating the intelligibility of song lyrics to provide an aid for language learners who listen to music as part of language immersion. We conducted a behavioral experiment to review how the intelligibility of lyrics in different genres of songs are perceived by human participants. We then developed a computational system to automatically estimate the intelligibility of lyrics in a given song. In our system, we proposed features to reflect different factors that affect the intelligibility of lyrics according to previous empirical studies. We used the proposed features along with standard audio features to train a model capable of estimating the intelligibility of lyrics (as low, moderate, or high intelligibility) with an AUC of 0.71. The study provides evidence that the proposed system has promising initial results, and draws attention to the problem of lyrics intelligibility, which has received little attention in terms of computational audio analysis and automatic evaluation.

7. REFERENCES

[1] Martha S Benolken and Charles E Swanson. The effect of pitch-related changes on the perception of sung vowels. *The Journal of the Acoustical Society of America*, 87(4):1781–1785, 1990.

[2] Ugo Cesari, Maurizio Iengo, and Pasqualina Apisa. Qualitative and quantitative measurement of the singing voice. *Folia Phoniatrica et Logopaedica*, 64(6):304–309, 2013.

[3] Lauren Collister and David Huron. Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review*, 3(3):109–125, 2–8.

[4] Nathaniel Condit-Schultz and David Huron. Catching the lyrics. *Music Perception: An Interdisciplinary Journal*, 32(5):470–483, 2015.

[5] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[6] Aihong Du, Chundan Lin, and Jingjing Wang. Effect of speech rate for sentences on speech intelligibility. In *Communication Problem-Solving (ICCP), 2014 IEEE International Conference on*, pages 233–236. IEEE, 2014.

[7] Harry Hollien, Ana Mendes-Schwartz, and Kenneth Nielsen. Perceptual confusions of high-pitched sung vowels. *Journal of Voice*, 14(2):287–298, 2000.

[8] Kristoffer Jensen and Tue Haste Andersen. Real-time beat estimationusing feature extraction. In *International Symposium on Computer Music Modeling and Retrieval*, pages 13–22. Springer, 2003.

[9] Randolph Johnson, David Huron, and Lauren Collister. Music and lyrics interaction and their influence on recognition of sung words: an investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming. *Empirical Musicology Review*, 9(1):2–20, 2014.

[10] Randolph B Johnson, David Huron, and Lauren Collister. Music and lyrics interactions and their influence on recognition of sung words: an investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming. *Empirical Musicology Review*, 9(1):2–20, 2013.

[11] Tung-an Kao and Rebecca Oxford. Learning language through music: A strategy for building inspiration and motivation. *System*, 43:114–120, 2014.

[12] Anne Kultti. Singing as language learning activity in multilingual toddler groups in preschool. *Early Child Development and Care*, 183(12):1955–1969, 2013.

[13] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. `https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`, 2007.

[14] Bernhard Lehner and Gerhard Widmer. Monaural blind source separation in the context of vocal detection. In *16th International Society for Music Information Retrieval Conference (ISMIR), At Malaga, Spain*, 2015.

[15] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. On the reduction of false positives in singing voice detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7480–7484. IEEE, 2014.

[16] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 53–56. IEEE, 2012.

[17] Paul Masri and Andrew Bateman. Improved modelling of attack transients in music analysis-resynthesis. In *ICMC*, 1996.

[18] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[19] Carmen Mora. Foreign language acquisition and melody singing. *ELT journal*, 54(2):146–152, 2000.

[20] Kazuma Mori and Makoto Iwanaga. Pleasure generated by sadness: Effect of sad lyrics on the emotions induced by happy music. *Psychology of Music*, 42(5), 2014.

[21] Tomoyasu Nakano. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proceedings of INTERSPEECH2006*, 2006.

[22] Joan Netten and Claude Germain. A new paradigm for the learning of a second or foreign language: the neurolinguistic approach. *Neuroeducation*, 1(1), 2012.

[23] Koichi Omori, Ashutosh Kacker, Linda Carroll, William Riley, and Stanley Blaugrund. Singing power ratio: quantitative evaluation of singing voice quality. *Journal of Voice*, 10(3):228–235, 1996.

[24] Aniruddh Patel. Language, music, syntax and the brain. *Nature Neuroscience*, 6(7):674–681, 2003.

[25] Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. PTR Prentice Hall, 1993.

[26] Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing*, 21(1):73–84, 2013.

[27] Mathieu Ramona, Gaël Richard, and Bertrand David. Vocal detection in music with support vector machines. In *Proc. ICASSP '08*, pages 1885–1888, March 31 - April 4 2008.

[28] Vishweshwara Rao, Chitralekha Gupta, and Preeti Rao. Context-aware features for singing voice detection in polyphonic music. In *International Workshop on Adaptive Multimedia Retrieval*, pages 43–57. Springer, 2011.

[29] Martın Rocamora and Perfecto Herrera. Comparing audio descriptors for singing voice detection in music audio files. In *Brazilian symposium on computer music, 11th. san pablo, brazil*, volume 26, page 27, 2007.

[30] Daniele Schon, Sylvain Moreno, Mireille Besson, Isabelle Peretz, and Regine Kolinsky. Songs as an aid for language acquisition. *Cognition*, 106(2):975–983, 2008.

[31] Robert Speer, Joshua Chin, Andrew Lin, Lance Nathan, and Sara Jewett. wordfreq: v1.5.1. `https://doi.org/10.5281/zenodo.61937`, September 2016.

[32] Valerie Trollinger. The brain in singing and language. *General Music Today*, 23(2), 2010.

[33] Xinxi Wang and Ye Wang. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 627–636. ACM, 2014.

[34] Christopher Watts, Kathryn Barnes-Burroughs, Julie Estis, and Debra Blanton. The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. *Journal of Voice*, 20(1):82–88, 2006.

[35] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

PRIMARY-AMBIENT EXTRACTION IN AUDIO SIGNALS USING ADAPTIVE WEIGHTING AND PRINCIPAL COMPONENT ANALYSIS

Karim M. Ibrahim
Nile University
k.magdy@nu.edu.eg

Mahmoud Allam
Nile University
mallam@nu.edu.eg

ABSTRACT

Most audio recordings are in the form of a 2-channel stereo recording while new playback sound systems make use of more loudspeakers that are designed to give a more spatial and surrounding atmosphere that is beyond the content of the stereo recording. Hence, it is essential to extract more spatial information from stereo recording in order to reach an enhanced upmixing techniques. One way is by extracting the primary/ambient sources. The problem of primary-ambient extraction (PAE) is a challenging problem where we want to decompose a signal into a primary (direct) and ambient (surrounding) source based on their spatial features. Several approaches have been used 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 to decompose the signal into primary and ambient sources using Principal Component Analysis (PCA) with an adaptive weighting based on the level of correlation between the two channels to overcome the problem of low ambient energy in PCA-based approaches.
Key words: Audio Source Separation, Primary/ambient Separation, Surrounding Sound Systems, Upmixing.

1. INTRODUCTION

Currently, most audio recordings are available as 2-channel stereo recordings. For a long time, this has been considered sufficient to give the listener a pleasant experience. However, with new sound systems that give a better sense of surrounding and enclosing atmosphere, older recordings fail to utilize the capabilities of these new systems. Thus, it is important to develop methods of extracting additional spatial information from these recordings to enhance the experience of listening to them: this process is called upmixing [1, 2]. One approach is applying audio source separation to extract the original sources from the mixture, which are then rendered for the new playback system [3]. An important distinction between the different audio sources that can be used as a base for separating the sources is the ability to localize the sound sources. Separating sources based on their directional and diffuse features can be used in upmixing to create an immersive feeling.

5.1 surround systems [4] are an example of a multi-channel sound system commonly used in home theaters that are often used to play stereo recordings. A practical method of upmixing the stereo sound to the 5.1 system is by separating the primary (localizable) and ambient (non-localizable) sources and playing the primary sources on the two front channels to recreate the direct sources as it was intended in the original recording while playing the ambient sources on all channels to give a better feeling of surround sound.

Such applications call for advanced audio source separation methods. Hence, such methods have increasingly gained attention in the research community. Audio source separation can generally be categorized into two main challenges: blind audio source separation (BASS), where the goal is to extract the different sound sources in the mix, and primary-ambient extraction (PAE), where the goal is to separate between primary (direct) sources and ambient (diffuse) sources.

Several approaches have been proposed to extract the primary and ambient sources from a mixed-down recording. A commonly used approach is using Principal Component Analysis (PCA) as in [5, 6], which is investigated in detail later in this paper as it is the basis for the proposed approach. A different approach for the problem is using the least square method to estimate the primary and ambient sources as proposed by Faller in [7] by minimizing the errors between the extracted signals and the original stereo input.

In Avendano's work [8], the approach is to calculate a band-wise inter-channel short-time coherence from the cross- and autocorrelation between the stereo channels which is then used as the basis for the estimation of a panning and ambiance index. In Kraft's approach [9], the proposed method is based on the mid-side decomposition of stereo signals where the two-channel recording is split into "mid" signal that captures the centered content of the recording and a "side" signal that captures the content panned to the left and right side.

The focus of this paper lies in developing a new technique for primary-ambient extraction in stereo signals and to introduce an evaluation method for PAE to compare between the different commonly used approaches and our new proposed method.

The paper is structured as follows: Section 2 explains the problem definition of audio source separation and primary ambient extraction, the possible application for these tech-

niques and the constraints for an ideal extraction.

Section 3 explains our proposed method to improve the separation based on Principal Component Analysis (PCA). Finally, Section 4 shows the evaluation between the proposed method and the previous methods from the literature.

1.1 Notation

The convention in this paper is to express signals in the time domain in lower case letter as x, while signals in the STFT domain are in upper case as X. Scalar variables are expressed in normal italic font as X while column vectors are expressed in bold italic font as $\boldsymbol{X}$ and matrices are expressed in bold non-italic font as $\mathbf{X}$.

Table 1 shows the commonly used symbols in this paper:

$\boldsymbol{x}$	Mixed stereo signal
$\boldsymbol{x}_l$, $\boldsymbol{x}_r$	left and right channels of a sound mixture
$\boldsymbol{p}_l$,$\boldsymbol{p}_r$	Left and right primary components
$\boldsymbol{a}_l$,$\boldsymbol{a}_r$	Left and Right ambient components
n	Discrete time index
m	Frequency index
k	Frame index
w_{pl},w_{pr}	weighting factor of the primary source
$\boldsymbol{v}$	Normalized unit vector of 1^{st} Principal component

Table 1: Symbols used in this paper

2. PRIMARY-AMBIENT EXTRACTION

One of the key characteristics in spatial audio is whether an audio source is localizable or not. A localizable source is perceived as coming from a certain direction and the listener can determine this direction, also called primary or directional source. A non-localizable source is perceived as a surrounding sound, coming from all around, also called an ambient or diffuse sound. Ambient sources usually describe the surrounding atmosphere of the recording . Methods for separating these two types of sources have been receiving increasing attention for applications such as upmixing [10, 11], multichannel format conversion and headphone reproduction [12, 13].

2.1 Signal model for PAE

When approaching the problem of primary-ambient extraction, we consider the input signal as a mix of two sources; a primary and an ambient source. In this paper, we only approach the problem of separating the mixture of a stereo signal.

Stereo recordings consist of two channels that contain both the primary and ambient sources mixed together and the goal is to separate them. The signals can be expressed as follows:

$$x_l[n] = p_l[n] + a_l[n] \quad (1)$$

$$x_r[n] = p_r[n] + a_r[n] \quad (2)$$

where x_l, x_r are the left and right channels of the stereo recording respectively, p_l, p_r are the primary component in each channel, a_l, a_r are the ambient component and n is the time index of the discrete signals.

Most PAE approaches are applied in the STFT domain as it is safer to assume there is only one primary source and one ambient source in each frequency-frame sub-band. The signals are expressed then in the form:

$$X_l[m,k] = P_l[m,k] + A_l[m,k] \quad (3)$$

$$X_r[m,k] = P_r[m,k] + A_r[m,k] \quad (4)$$

where m, k are the frame and frequency index respectively.

2.2 Sound localization and human auditory system

To be able to precisely separate the primary and ambient sources, it is necessary to understand how the human auditory system works and how it determines the location of a sound source and then use the same characteristics in the separation process.

The human auditory system uses several cues to localize a sound source, including inter-channel time difference (ICTD), also referred to as inter-aural time difference (ITD), inter-channel level difference (ICLD), also referred to as inte-aural level difference (ILD), spectral information and correlation analysis [14].

A comparison between the two channels should be sufficient to extract the directional information of an audio source. The correlation between the two channels plays a significant role in determining the location of the source, i.e., an ambient source shows no correlation between the two channels, making it impossible for the human auditory system to determine the direction of the sound. Hence, calculating the correlation between the two channels is usually a necessary step in extracting the primary and ambient sources.

2.3 PAE applications: upmixing to 5.1 systems

A common application for PAE is upmixing from n to m channels, where $m > n$. Here, we explain how to use PAE in upmixing to one of the commonly used systems, the 5.1 surround system. By separating the primary and ambient sources using one of the PAE methods, the extracted sources are re-panned in a way that the left primary sources $p_l[n]$ are played on the front left and center channels, $x_{lf}[n]$ and $x_c[n]$ while the right primary sources are played on the front right and center channels $x_{rf}[n]$ and $x_c[n]$ and the ambient sources are played throughout the five speakers. This way, the directionality of the primary sources are kept as originally intended while the surrounding sound is enhanced by the ambient sources. Figure 1 shows the block diagram of the upmixing technique.

2.4 PAE assumptions

To accurately separate between the primary and ambient components, we need to define the constraints that achieve the right separation. By definition, the primary sources are localizable while the ambient sources are non-localizable.

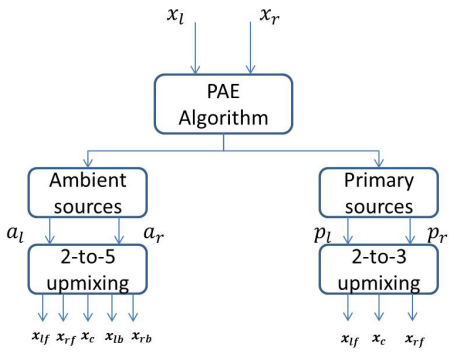

Figure 1: Block diagram of the stereo to 5.1 upmixing using PAE

To find a mathematical representation for this definition, we need to review the sound localizing process in the human auditory system mentioned in section 2.2. The key characteristic in localizing the sound sources is the correlation between the two signals reaching the left and right ears. In the case of a complete non-localizable diffuse source, the two signals are expected to be orthogonal in a way that the brain fails to detect any similarity between the left and right signals to extract location information. Similarly, primary sources are expected to be partially or fully correlated. Based on the representation of the stereo signal in equation (3) and assuming that the left and right primary components are P_l, P_r respectively, where P_l, P_r are vectors of adjacent STFT frames, Similarly the left and right ambient components are A_l, A_r, ω is the scaling factor between the primary components in the two channels due to ICLD and $\boldsymbol{A}^H$ is the Hermitian transpose of the vector $\boldsymbol{A}$, these constraints are defined according to [5] as:

1. The primary components are correlated

$$\boldsymbol{P}_l = \omega \boldsymbol{P}_r \tag{5}$$

2. The ambient components are orthogonal (fully uncorrelated)

$$\boldsymbol{A}_l^H \boldsymbol{A}_r = 0 \tag{6}$$

3. The ambient and primary components are orthogonal to each other

$$\boldsymbol{P}_l^H \boldsymbol{A}_l = 0 \qquad \boldsymbol{P}_r^H \boldsymbol{A}_r = 0 \tag{7}$$

4. The two ambient components have almost the same energy level

$$\boldsymbol{A}_l^H \boldsymbol{A}_l \approx \boldsymbol{A}_r^H \boldsymbol{A}_r \tag{8}$$

Figure 2 shows the assumed constraints between the different components.

2.5 PCA-Based PAE

Many of the approaches of PAE are based on the Principal Component Analysis (PCA) as in [5,6,15–18]. PCA is widely used since the common signal model assumes that the stereo signal is composed of primary sources that are

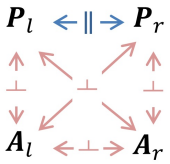

Figure 2: Constraints on the primary and ambient components

highly correlated and ambient diffuse sources. It is suitable to use a decomposition method such as PCA to extract the correlated primary sources and to assume the ambient sources are the residuals. The work in [15] is also based on the PCA but with an important modification, it takes into consideration the Inter-Channel Time Difference (ICTD) by using a time-shifting technique to improve the extraction of the primary sources.

One major drawback of methods based on PCA is the assumption that there is always a primary source in each frequency-frame sub-band and that it is never too weak. This is evident from the extraction of the primary sources as the first principal component. In case of absence of any primary sources, the method would still assign the first principal component, the one with the highest energy, to the primary source, which clearly produces a significant error in this particular case.

3. IMPROVING PCA-BASED APPROACH

As described in Section 2.5, the PCA-based approach has a number of drawbacks that impairs its accuracy. The solution we propose is to add an adaptive weighting to increase the amount of energy the ambient signal. The concept of adaptive weighting in PCA was previously introduced by Goodwin [19] with a different weighting scheme. The weighting we propose is based on the relation between the two channels of the signal in a way that supports the ambient extraction by detecting the level of presence of the primary sources. One way to do this is by considering the second dominant eigenvalue and comparing its value to the dominant eigenvalue. In the case of high correlation, the first (dominant) eigenvalue will be considerably larger than the second eigenvalue. In this case it would be safe to decompose the signal into primary and ambient components. However, in the case of having a more dominant ambient source, the ratio between the first and second eigenvalues will be relatively small.

The PAE using our weighting scheme is applied as follows:

1. We start with the original 2-channel signals, $x_l[n]$ and $x_r[n]$. We apply the STFT on the signals to get $X_l[m,k]$ and $X_r[m,k]$, where m is the frame index and k is the frequency index. We calculated the STFT using $\frac{3}{4}$ overlapping Hamming windows of Length 4096 samples, corresponding to a duration of 92.8 milliseconds at a sampling frequency of 44.1 kHz.

2. For each frequency-frame bin we define a vector with

the STFT values of the M adjacent frames:

$$\boldsymbol{X}_{l,r}[m,k] = \begin{bmatrix} X_{l,r}[m-M,k] \\ \vdots \\ X_{l,r}[m+M,k] \end{bmatrix} \quad (9)$$

For brevity, the index $[m,k]$ is dropped in the following equations.

3. The decomposition is then applied per frame-frequency index to extract the primary and ambient components in each frame-frequency using these two vectors as shown in Figure 3.

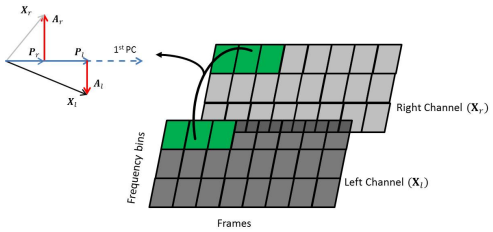

Figure 3: PAE using PCA

4. Using the Eigendecomposition on the covariance matrix $\mathbf{C}$ of the two vectors $\boldsymbol{X}_l, \boldsymbol{X}_r$, we get the normalized dominant Eigenvector $\boldsymbol{V}$ and the first two dominant Eigenvalues λ_1, λ_2.

5. Next we calculate the weighting factor ω based on the ratio between the two eigenvalues. The primary weights are defined as:

$$\omega = 1 - \frac{\lambda_2}{\lambda_1} \quad (10)$$

This weighting ensures that in cases of low correlation between the two channels, a higher weight is given to the ambient component.

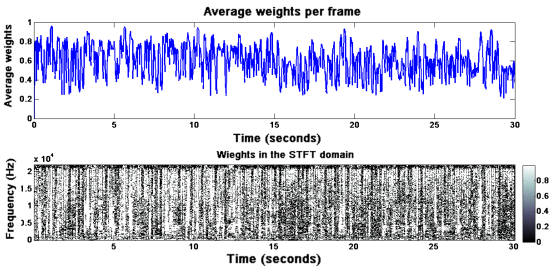

Figure 4: Sample of the weights extracted from an audio file

6. In order to enhance the ambient extraction, we define a threshold θ. The goal is to detect the cases where there is no strong presence of a primary source, so all the content is put into the ambient component. The primary component is still weighted by ω in case of passing the threshold to support extracting the ambient component.

$$\boldsymbol{P}_{l,r} = \begin{cases} \omega(\boldsymbol{V}^H \boldsymbol{X}_{l,r})\boldsymbol{V}, & \omega > \theta \\ 0, & \omega < \theta \end{cases} \quad (11)$$

$$\boldsymbol{A}_{l,r} = \boldsymbol{X}_{l,r} - \boldsymbol{P}_{l,r} \quad (12)$$

where P_l, P_r, A_l, A_r are the primary and ambient components of the right and left channel respectively. Figure 4 shows an example of the weights extracted from an audio file.

7. Finally, two schemes can be used to extract the information from each frame; either to merge the extracted vectors by averaging them or to take out the center point of each vector as shown in figure 5. However, the results of both schemes are very similar, so we can use the "take center" scheme to reduce the computation.

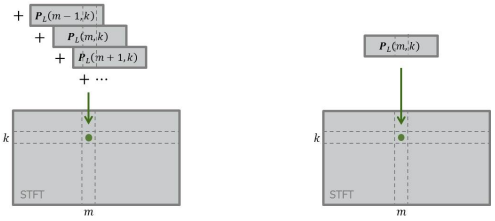

Figure 5: Different schemes of merging the output vectors

4. EVALUATION

Our objective evaluation is based on the work done in [20] which is intended to evaluate blind audio source separation (BASS). It can be used for primary-ambient separation, as well, by assuming that the mixture of sources is made out of only two sources, an ambient and a primary one. Ideally the extraction methods should output two sources that are identical to the originals ones. However, due to the limitations of the extraction methods, there is interference between the two sources.

In the following we would like to compare the following approaches:

1. PCA-based PAE without weighting, referred to as "PCA without weighting".

2. The weighted PCA method by Goodwin in [5, 6]. Referred to as (PCA Goodwin).

3. The extraction method by Avendano and Jot in [8]. referred to as (Avendano)

4. The modified PCA method with the weighting scheme proposed in this paper with two different threshold values: $\theta = 0.5$ and $\theta = 0.9$.

The evaluation was performed using two databases, one is made out of all ambient sources, consisting of strong ambient sources as sounds of crowd, forest, rain and echoes, and the second is made of all primary sources, consisting of strong primary sources as vocal recordings, solo instruments and dialogs. Each of the two data sets consist of 40 different recordings that are mixed together to compose 40 mixed recordings. We used the Matlab toolbox "BSS Eval" [21] for calculating the errors. The evaluation is as follows:

1. Mixing one ambient source with one primary source after normalizing the two of them, by ensuring the highest energy level of the two sources is the same, so no source would be more prominent than the other.

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 using BSS Eval.

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.

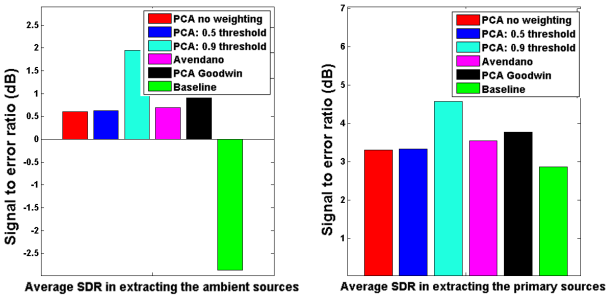

Figure 6: Average SDR in primary and ambient extraction

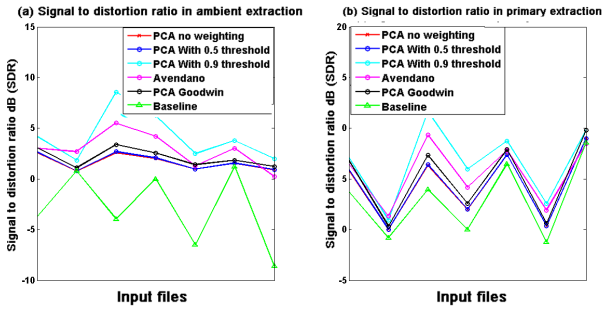

Figure 7: SDR values for a sample of five mixtures

Figure 6 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 proposed weighting shows an improvement in the separation over the other methods. We find that using a higher threshold of 0.9 gives much better separation than using a lower threshold or no threshold. This shows how the weighting improves the accuracy of extraction over both the original PCA and the weighted PCA introduced by Goodwin in [19].

Figure 7 shows the exact SDR values of a sample of five mixtures with comparison to the baseline in both the primary and ambient extraction. We find that all the methods improve clearly over the baseline without separation. In general the SDR values for the primary extraction is higher than the ambient because the primary sources tend to have higher energy.

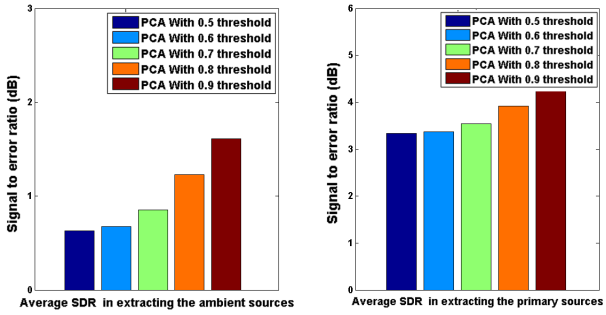

Figure 8: Average SDR for different thresholds

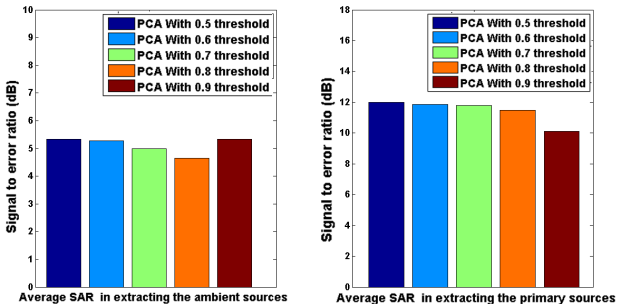

Figure 9: Average SAR for different thresholds

Figure 8 shows how using different thresholds affects the extraction quality. Using higher thresholds gives higher accuracy in the separation, however, extreme weights result in higher distortion caused by the artifacts in the separation process, especially in the extracted primary sources, as shown in Figure 9. Hence, there is a trade-off between sharp separation and artifact distortion. Typically, a threshold in the range $\theta \in [0.6, 0.8]$ would give a proper trade-off between separation quality and artifact distortion.

5. CONCLUSIONS

Separating the primary and ambient sources from an audio mixture shows potential for applications including upmixing an audio recording. In this paper, we explained the need for this separation technique and proper ways of using it in upmixing techniques. We presented a method of extracting the sources using an adaptive Principal Component Analysis (PCA) to solve the common problem of the dominant primary source. The adaptive weighting tests the level of presence of primary sources and ensures to give a proportional weight to both of the sources based on this estimate. The method shows higher separation quality compared to the classic PCA-based separation methods and other methods from the literature. However, this method still shows correlation between the two ambient components leaving room for further improvement in future work. Future work could also include a subjective

evaluation by performing listening test with the different separation methods to ensure the user's experience coincide with the results of the objective evaluation.

6. REFERENCES

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

[2] M. R. Bai and G.-Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *J. Consumer Electronics*, vol. 53, no. 3, pp. 1011–1019, 2007.

[3] D. Fitzgerald, "Upmixing from mono-a source separation approach," in *Proc. Int. Conf. Digital Signal Processing (DSP), 2011*. IEEE, 2011, pp. 1–7.

[4] B. Xie, "Signal mixing for a 5.1-channel surround sound system'analysis and experiment," *J. Audio Engineering Society*, vol. 49, no. 4, pp. 263–274, 2001.

[5] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 2007, pp. I–9.

[6] M. M. Goodwin, "Geometric signal decompositions for spatial audio enhancement," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 409–412.

[7] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Engineering Society*, vol. 54, no. 11, pp. 1051–1064, 2006.

[8] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Engineering Society*, vol. 52, no. 7/8, pp. 740–749, 2004.

[9] S. Kraft and U. Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, 2015.

[10] M. R. Bai and G.-Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *J. Consumer Electronics*, vol. 53, no. 3, pp. 1011–1019, 2007.

[11] C. Faller and J. Breebaart, "Binaural reproduction of stereo signals using upmixing and diffuse rendering," in *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[12] J. Breebaart and E. Schuijers, "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *J. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1503–1511, 2008.

[13] W.-S. Gan, E.-L. Tan, and S. M. Kuo, "Audio projection," *J. Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 43–57, 2011.

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

[15] J. He, E.-L. Tan, and W.-S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 2013, pp. 266–270.

[16] S.-W. Jeon, D. Hyun, J. Seo, Y.-C. Park, and D.-H. Youn, "Enhancement of principal to ambient energy ratio for pca-based parametric audio coding," in *Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 385–388.

[17] J. Merimaa, M. M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[18] S. Dong, R. Hu, W. Tu, X. Zheng, J. Jiang, and S. Wang, "Enhanced principal component using polar coordinate pca for stereo audio coding," in *Proc. Int. Conf. Multimedia and Expo (ICME)*. IEEE, 2012, pp. 628–633.

[19] M. M. Goodwin, "Adaptive primary-ambient decomposition of audio signals," Jun. 19 2012, uS Patent 8,204,237.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *J. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.