Audio Engineering Society

Convention Express Paper 62

Presented at the 154th Convention
2023 May 13–15, Espoo, Helsinki, Finland

Deep Learning Based Voice Extraction and Primary-Ambience Decomposition for Stereo to Surround Upmixing

Ricardo Thaddeus Páez-Amaro[1], Carlos Tejeda-Ocampo[1], Ema Souza-Blanes[2], Sunil Bharitkar[2], and Luis Madrid-Herrera[1]

[1]*Samsung Research Tijuana*
[2]*Samsung Research America*

Correspondence should be addressed to Thaddeus Páez (t.paez@samsung.com)

ABSTRACT

Surround systems have gained popularity in home entertainment despite the fact that most of the cinematic content is delivered in two-channel stereo format. Although there are several upmixing options, it has proven challenging to deliver an upmixed signal that approximates the original directionality and timbre intended by the mixing artist. The aim of this work is to design a two-to-five channels upmixer using a novel upmixing strategy combining voice extraction and primary-ambience decomposition. Results from a modified-MUSHRA test show that our proposed upmixer outperforms established alternatives for cinematic upmixing in perceived spatial and timbral quality.

1 Introduction

Today, multichannel surround home theaters have become more accessible to consumers. However, most audiovisual content remains in stereo format. Since playing stereo content in surround systems does not offer the best possible listening experience, upmixing techniques have been used to derive signals in surround formats (e.g. 5.1, 7.1, 7.1.4) from an original 2-channel mix. Upmixing is the process where audio content of m channels is mapped into n channels, where $n > m$. These n-channels should be able to be played in a surround speaker setup and provide a better immersive experience to the listener than plain stereo.

Various upmixing methods have been proposed: the passive matrix [1], least squares estimates [2], subjectively tuned mapping functions [3, 4], Principal Component Analysis (PCA) [5, 6], Normalized Least-Mean-Square (NLMS) adaptive filter in frequency domain and time domain Least-Mean-Square (LMS) filter [7], mid-side (M-S) decomposition [8], neural networks (NN) [9, 10] and style transfer [11]. Nevertheless, these methods can present phasiness, especially when moving outside the sweet spot or when playing back a subset of the five speakers [4], or do not perform well when the input channels are already uncorrelated, or

do not sound that natural [12], or are designed for a particular type of content, e.g. music. [11].

This paper proposes the Voice-Primary-Ambience Extraction Upmixing (VPA) method. VPA focuses on upmixing from two to five channels. It consists of three main blocks: vocal extraction, primary-ambience decomposition, and upmix rendering. Vocals and ambience extraction is performed using the OpenUnmix (UMX) [13] and Equal-Levels Ambience Extraction (ELAE) [12] algorithms respectively. UMX is an algorithm that extracts vocals using a three layer Bidirectional Long Short-Term Memory (BiLSTM) network [13]. UMX is used because is open source, easy to use and still yields state-of-the-art results. ELAE extracts the ambience assuming that the environment has the same level on the left and right channels in typical stereo recordings. Finally, VPA upmix rendering consists of processing and transferring information obtained through UMX and ELAE to the different channels. This rendering process will be explained in more detail in Section 2.3.

The proposed VPA algorithm is subjectively evaluated through a modified MUSHRA test [14, 15]. The results obtained by these evaluation methods demonstrate that VPA is robust and reliable, outperforming both naive approaches like mid-side upmixing (M-S) and a commercial alternative.

This paper is organized as follows. Section 2 describes the methods used for this research. Section 3 presents the results obtained by the subjective tests. Section 4 discusses the results obtained by the proposed algorithm. Finally, Section 5 summarizes the conclusions of this work.

2 Methods

2.1 Voice Extraction

Unmixing refers to the process of separating the different sources which comprise a signal. The nature of the audio sources present will vary depending on the type of audio signal being upmixed. In music the common sources are predictable to a certain extent: vocals, guitar, keyboard, bass, drums, among others. However, in cinematic content, there could be an unpredictable number of sources of different kinds. This makes unfeasible to implement a broader sound separation approach for cinematic content upmixing.

The most common approach to do unmixing is by finding source patterns in the mix spectrogram and extracting it through a mask. However, there are different methods to achieve this, such as harmonic-percussion separation [16], non-negative matrix factorization (NMF) [17] (as cited in [18]), or neural networks [13, 19, 20].

OpenUnmix (UMX) [13] is a Deep Learning model, trained for source separation task in a musical context. For this paper, we used the *umxl vocals* model with the pre-trained weights provided in [21]. Although the model was trained to extract singing voices it performs well extracting speech from cinematic content, however, the vocal reverberation is not included in the extracted speech signal but is found in the residual signal in both cinematic and musical content cases. The core of the UMX architecture is a 3-layer BiLSTM. It takes as input the STFT spectrogram of the mix, crops it to 16 kHz, passes it through a fully connected layer, then through the BiLSTM, and two more fully connected layers, including additionally a skip connection right before and after the BiLSTM. Finally, it reshapes the output to match the original STFT shape and outputs a mask, which will be applied to the original spectrogram to perform the actual source extraction. UMX offers a residual estimation of what is left after the desired source was extracted from the mix.

2.2 Ambience extraction

VPA uses the Equal-Levels Ambience Extraction (ELAE) algorithm [12]. ELAE is based on the following assumptions: (i) An input signal is the result of adding up a primary (directional) component and ambience; (ii) in a stereo signal, the primary components are uncorrelated with their ambience, and the ambience signals are uncorrelated with each other; (iii) the correlation coefficient of the primary components is 1; (iv) ambience levels in both channels are equal; (v) it is possible to extract the ambience through a mask.

Using the above assumptions and the physical constraint that the total ambience energy has to be lower than or equal to the total energy it is possible to find the masks as a function of the channels' cross-correlation and auto-correlations [12].

2.3 Upmixing

VPA consists of three main blocks: voice extraction, ambience extraction and upmix rendering. A diagram of the whole process is depicted in Fig. 1.

The first block comprises of the pretrained *vocals* model of UMX [22] as source extractor. It receives the stereo downmix and produces a 4-channel audio, i.e., the concatenation of the extracted voice in stereo ($[V_L, V_R]$) with the residual also in stereo ($[U_L, U_R]$).

For the first block, let s be the stereo input signal with s_L and s_R its left and right channels, respectively.

$$[V,U] = UMX(s) \tag{1}$$

Where V is the extracted voice, and U is the residual of s after removing V.

The second block is the Primary-Ambience decomposition, which is done just over the residual U using ELAE.

$$[P,A] = ELAE(U) \tag{2}$$

where P contains the primary component of U and A contains the ambience of the residual U.

Next, the upmix rendering block. One step before obtaining the upmixed signal $\hat{s}$, the pre-upmixed channels $\bar{s}_{\{L,R,C,L_s,R_s\}}$ are generated as follows. $\bar{s}_L$ is the mix of V_L(-48 dB) and P_L(-1 dB). Likewise, $\bar{s}_R$ is the mix of V_R(-48 dB) and P_R(-1 dB). Then, A_L(+12 dB) and A_R(+12 dB) are decorrelated through a 64th-order all-pass filter to get $\bar{s}_{Ls}$ and $\bar{s}_{Rs}$, respectively. Decorrelation is applied to broaden the sound and extend the surrounding perception according to [25]. Center channel $\bar{s}_C$ is the downmix of stereo voice V(-3 dB) and the stereo primary component P(-48 dB). Finally, a 2 dB bass cut is applied to frontal channels ($\bar{s}_L$, $\bar{s}_R$, $\bar{s}_C$) and a 2 dB bass boost to the rear channels ($\bar{s}_{Ls}$, $\bar{s}_{Rs}$), using a low-pass shelving filter with slope of 0.8 and half-gain frequency at 250 Hz.

2.3.1 Real-Time Approach

In order for VPA to be implemented in a consumer application it needs to be performed in real time. To achieve this, we propose a windowed approach, where small chunks of the audio are processed in overlapping slices. We chose a window size $W = 4096$ with an overlap $O = 512$ samples. Nevertheless, UXM was trained using STFT windows with 4096 samples and overlap

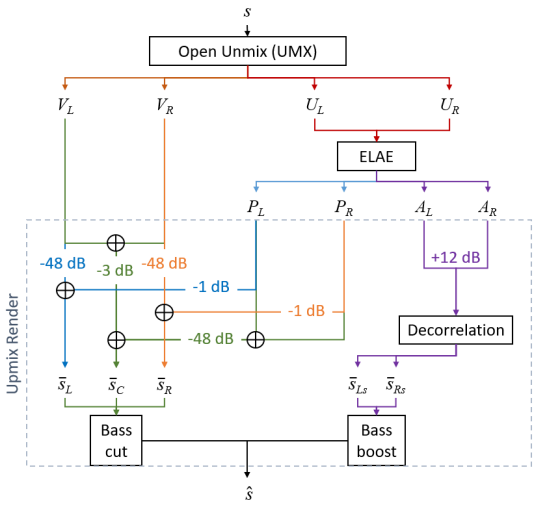

Fig. 1: Flowchart diagram of our proposed stereo-to-surround upmixing method.

of 3072 samples, so we kept that configuration in the internal UMX block; and for the ELAE's internal STFT we used a 128-sample window with 96 overlapping samples. To address the border artifacts, inherent to the STFT process and due to the rears' decorrelation, we take out the last $cE = 96$ samples of each window and the first $cS = 416$ samples of the next window before concatenating them. The pseudocode for this approach is shown in Algorithm 1, where N is the total number of processed windows, $\hat{s}$ is the upmixed signal corresponding to the current window, and *upmix* is the final output with the complete upmixed signal.

In Table 1 the latencies per block are shown. The presented computing times were measured when running in a Core i7-7700HQ @ 2.8 GHz CPU with 16 GB of RAM. The sampling frequency of the audio used for testing was 48 kHz, so each window represents 85 ms. The entire algorithm takes on average 55.76 ± 4.16 ms per window, lower than the actual time each window represents. Note that the reported latencies are just the time it takes to process a single window, without considering any overlap. To take it into account we note that the chosen overlap O represents 12.5% of W, which corresponds to a proportional delay, keeping the latency within the 85 ms window even when overlap is considered. This means it is feasible for VPA to work in a real-time environment.

In our setup we used the PyTorch version of UMX but

Algorithm 1: Windowed VPA

Input: s
Output: upmix
Require : $s \geq 4096$
$W \leftarrow 4096$;
$O \leftarrow 512$;
$cS \leftarrow 416$;
$cE \leftarrow 96$;
$N \leftarrow (len(s) - O)/(W - O)$;
for $n \leftarrow 1$ **to** N **do**
  $startIdx \leftarrow (n-1)(W-O)+1$;
  $endIdx \leftarrow startIdx + W - 1$;
  $\hat{s} \leftarrow VPA(s[startIdx : endIdx])$;
  if *n is 1* **then**
    $upmix[startIdx : endIdx] \leftarrow \hat{s}$;
  else
    $upmix[startIdx + cS : endIdx - cE] \leftarrow$;
    $\hat{s}[1 + cS : end - cE]$;
  end
end

the rest of VPA was coded in MATLAB. For that reason we did the voice extraction first in one run and then passed the output to the real-time version of ELAE and the upmix render blocks in a second run. The upmixed audio was normalized to -23 LUFS using the MATLAB's *integratedLoudness* function. Those final output files were then used for the subjective tests.

2.4 Tests

2.4.1 Benchmark

We present a comparison benchmark with two other common upmixing methods:

- Dolby Surround: The Dolby proprietary Surround Upmixer analyzes and processes multiple perceptually spaced frequency bands to separate steered and diffused sources, then positions each individually [23]. The stereo files were sent to a Marantz AV7706 pre-amplifier, with the Dolby Digital Surround upmixer setting. Each of the 12 channels of the resulting audio hardware output was converted to digital by way of an RME M-32 AD, and sent to a laptop running Nuendo 12 through an AVB network. All channels were then combined to obtain the Dolby Surround upmixed audio files.
- M-S: A naive approach where the center channel is the sum of the left and right channels, surround left is right minus left, and surround right is left minus right. A treble cut is applied to the rear channels at 2 kHz.

Table 1: Mean computing time for each block of VPA.

Block	Mean latency (ms/window)
Voice Extraction	42.73 ± 3.67
Primary-Ambience Extraction	8.85 ± 0.29
Upmix render	4.18 ± 0.2

2.4.2 Subjective tests

To test the performance of our proposed VPA upmixer against Dolby and M-S methods, we performed a modified version of the MUSHRA-test [14]. Five cinematic audio samples were downmixed to stereo, then upmixed back to five channels using VPA, Dolby and M-S. The original 5.1 version of the audio was used as both reference and hidden reference. The stereo downmix processed with a low-pass filter was employed as a low anchor. For the rest of the paper, the benchmarking conditions Dolby and M-S will be referred to as systems A and B, in no particular order.

[24] proposes that the original 5.1 quality can be assumed to be an "optimal" reference or "high anchor" for comparison with upmixing methods. While perfect reconstruction of the original signal is not feasible from a blind upmixing standpoint, it is reasonable to assume that a "good" upmixer would approximate the "optimal" spatial and timbral characteristic of the original version. This is specially true for cinematic content, where sound sources are positioned in the audio mix to correspond with the position of the sources in the screen. An upmixer for cinematic content should not deviate too drastically from the spatiality of the original content, else coherence between audio and image can be lost. For this reason, the original 5.1 version was chosen as a reference for the modified MUSHRA subjective test.

This resulted in a test where each trial included hidden upmix to be compared to the original reference, and rated over a scale ranging from 0 to 100. Listeners were invited to rate the spatial quality and the timbral

quality in two separate test sessions. The order of the sessions was manually distributed across the assessor panel.

2.4.3 Setup

Fig. 2: Listening room

Listening tests took place in a 7 m (L) x 5.33 m (W) x 3.05 m (H) listening room (Figure 2) supporting 5.1 channel-based playback. The loudspeakers' and listeners' positions were based on the ITU-R BS.2051-2 [25] standard. Loudspeakers were level-matched at the listeners's position and met the ITU-R BS.1116-3 [26] specification in terms of room response curve within the 50 Hz -16 kHz frequency range.

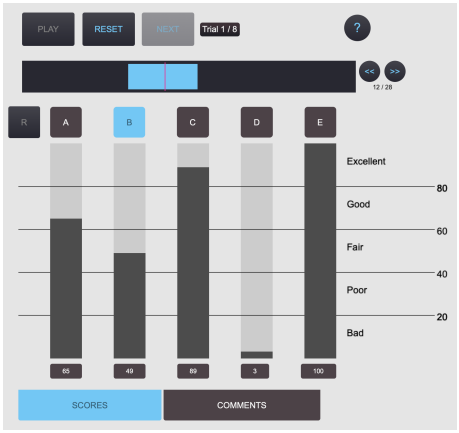

Fig. 3: Test user interface

A tablet interface using Max/MSP (Figure 3) provided the assessors with control over selection and playback of the test audio; as well as a rating module and the option to leave comments. A segment looping function enabled the assessors to focus on artifacts in restricted sections of the audio clips. The listening test software was implemented using a customised Max/MSP program to achieve a double-blind system, allowing for randomization of audio sample playback order and testing item mapping order on each trial.

3 Results

A panel of 14 assessors, including 11 experienced and 3 naive listeners, took part in the experiment.

After screening the assessors panel for outliers, the data were checked for ANOVA assumptions (normal distribution of the residuals and homoscedasticity).

For each test session, a two-way ANOVA was performed to analyze the effect of the condition (upmixing method) and audio sample on the scores collected for each attribute. It revealed that in both tests, there was a significant effect of the condition ($p < 2.2e^{-16}$ for both tests) and of the sample ($p = 0.007$ for spatial quality, $p = 0.005$ for timbral quality).

No statistically significant interaction between the effects of the two factors was found ($F(16,4) = 1.198$, $p = 0.268$ for spatial quality; ($F(16,4) = 1.57$, $p = 0.077$ for timbral quality).

No upmixing method reached the spatial or timbral quality of the reference sample, according to the assessors panel, with the hidden reference scoring higher on average than every other condition (See Figure 4 and Figure 5). VPA was on average rated higher than conditions other than the hidden reference, and System B was rated lower on average (cf mean scores in Table 2).

Table 2: Mean scores, per test attribute and upmixing method

Mean Score	Hidden Ref.	VPA	System A	System B
Spatial Quality	97.6	61.1	41.6	38.5
Timbral Quality	97.5	59.1	54.8	33.4

Tukey's HSD Test for multiple comparisons showed the mean scores were significantly different between all pairs of conditions in both tests, with two exceptions:

- scores given to System A and System B conditions were found not significantly different for Spatial Quality ($p = 0.85$, $95\%C.I. = [-10.7, 4.91]$);

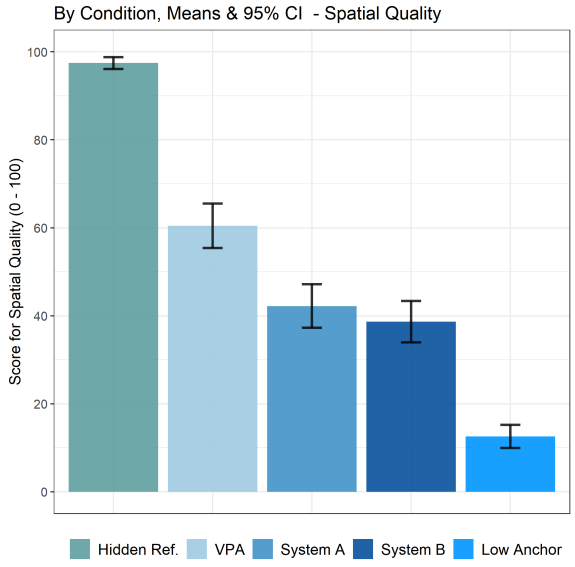

Fig. 4: Results for the Spatial Quality listening test, per condition

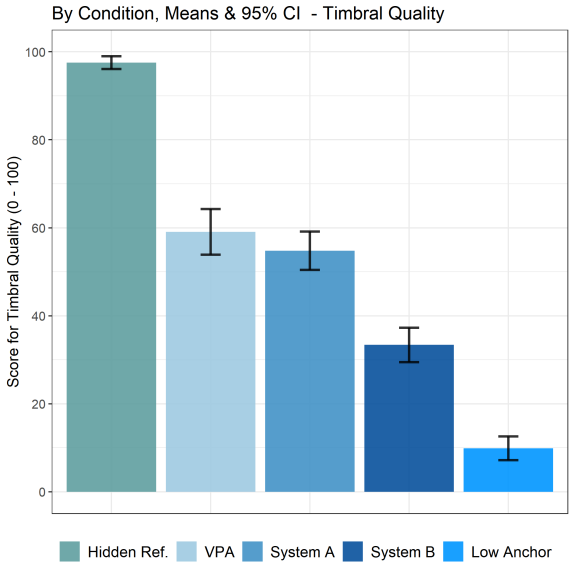

Fig. 5: Results for the Timbral Quality listening test, per condition

- scores given to VPA and System A conditions were found not significantly different for Timbral Quality ($p = 0.49$, $95\%C.I. = [-11.6, 3.02]$).

Comments left by the listeners for spatial quality revealed that the System B and System A conditions were similarly commented on as narrow, centered and lacking spatial definition. System B was also found distant, while frequent spatial errors were reported for System A, with too much content being displaced to the rear. Although several assessors reported that the VPA condition was also centered, the sound stage was described as the most natural and coherent.

Regarding the timbral quality, VPA, System B and System A conditions were comparably reported as muffled compared to the reference. System A was additionally described as tubby and hollow, with reduced voices. System B was repeatedly commented on as heavy and distorted in the midrange.

4 Discussion

One advantage of sending the voice to a true center channel is that it tends to be more stable than the phantom source generated through L and R channels, as stated in [24], leading to a robust 3D spatial rendering even when the listener is not exactly at the sweet spot.

In the case of cinematic mixes, the voice content tends to be panned to the center channel. Although this would not always be the case, the present work suggests that unmixing and panning voice to the center channel can be a generally good strategy for upmixing. Dialogue is a very important part of the cinematic experience, isolating voice before upmixing can prevent artifacts generated by the primary-ambience extraction processes that could impair intelligibility.

Voice unmixing could also have other advantages. For example, it could be implemented alongside dialogue enhancement processing to further upgrade the listening experience. Future work should also address cases where voices in the original mix are coming from a direction other than the center channel, and develop strategies to consistently identify such cases and steer the voice signal to the correct direction.

For subjective testing of upmixing methods, it is important to evaluate both spatial and timbral qualities separately. An upmixer could have good performance in spatial quality while underperforming in timbre, or vice versa. Since perfect reconstruction of the original signal is not feasible, objective similarity metrics are of

limited usefulness for evaluate upmixing models. For this reason, subjective testing is most important to assess upmixing quality and inform further development.

5 Summary

We presented a novel strategy for upmixing cinematic content which combines voice unmixing with primary-ambience extraction. This new approach is feasible to be implemented with real-time processing. Subjective tests indicate this approach can outperform commercial alternatives regarding spatial fidelity to the original audio. Future work should focus on improving timbral quality and voice directionality.

6 Acknowledgements

This research was supported by Samsung Research America and Samsung Research Tijuana. The authors would like to express their gratitude to the participants of the listening experiments.

References

[1] Dressler, R., "Dolby Surround Pro Logic decoder principles of operation," *Dolby White paper*, 2000.

[2] Faller, C., "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, 54(11), pp. 1051–1064, 2006.

[3] Avendano, C. and Jot, J.-M., "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. II–1957, IEEE, 2002.

[4] Avendano, C. and Jot, J.-M., "Frequency domain techniques for stereo to multichannel upmix," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Audio Engineering Society, 2002.

[5] Baek, Y.-H., Jeon, S.-W., Park, Y.-c., and Lee, S., "Efficient primary-ambient decomposition algorithm for audio upmix," in *Audio Engineering Society Convention 133*, Audio Engineering Society, 2012.

[6] Goodwin, M. M. and Jot, J.-M., "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, pp. I–9, IEEE, 2007.

[7] Irwan, R. and Aarts, R. M., "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, 50(11), pp. 914–926, 2002.

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

[9] Choi, J. and Chang, J.-H., "Exploiting Deep Neural Networks for Two-to-Five Channel Surround Decoder," *Journal of the Audio Engineering Society*, 68(12), pp. 938–949, 2021.

[10] Park, S. Y., Chun, C. J., and Kim, H. K., "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 377–380, IEEE, 2016.

[11] Yang, H., Wager, S., Russell, S., Luo, M., Kim, M., and Kim, W., "Upmixing via style transfer: a variational autoencoder for disentangling spatial images and musical content," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 426–430, IEEE, 2022.

[12] Merimaa, J., Goodwin, M., and Jot, J.-M., "Correlation-Based Ambience Extraction from Stereo Recordings," in *Audio Engineering Society 123rd Convention*, Audio Engineering Society, 2007.

[13] Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y., "Open-Unmix - A Reference Implementation for Music Source Separation," *Journal of Open Source Software*, 4(11), pp. 1667–1673, 2019.

[14] ITU-R BS.1534-2, Standard, International Telecommunication Union, Geneva, CH, 2014.

[15] souza blanes, e., tejeda ocampo, c., wang, c., and bharitkar, s., "bitrate requirements for opus with

first, second and third order ambisonics reproduced in 5.1 and 7.1.4," *journal of the audio engineering society*, 2022.

[16] Fitzgerald, D., "Harmonic/Percussive Separation using Median Filtering," *13th International Conference on Digital Audio Effects (DAFX10)*, 2010.

[17] Virtanen, T., "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, 15(3), pp. 1066–1074, 2007.

[18] Ozerov, A., Févotte, C., and Vincent, E., "An introduction to multichannel NMF for audio source separation," *Audio Source Separation, Springer, In press. ffhal-01631187v1*, 2018.

[19] Wang, Y., Stoller, D., Bittner, R. M., and Bello, J. P., "Few-Shot Musical Source Separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121–125, 2022.

[20] Manilow, E., O'Reilly, P., Seetharaman, P., and Pardo, B., "Source Separation By Steering Pretrained Music Models," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 126–130, 2022.

[21] Stöter, F.-R. and Liutkus, A., "Open-Unmix-Pytorch UMX-L," 2021, doi:10.5281/zenodo.5069601, [Dataset].

[22] Stöter, F.-R. and Liutkus, A., "Open-Unmix for PyTorch," 2019, `https://github.com/sigsep/open-unmix-pytorch`.

[23] Vinton, M., McGrath, D., Robinson, C., and Brown, P., "Next Generation Surround Decoding and Upmixing for Consumer and Professional Applications," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*, Audio Engineering Society, 2015.

[24] Chétry, N., Pallone, G., Emerit, M., and Virette, D., "A Discussion about Subjective Methods for Evaluating Blind Upmix Algorithms," *Audio Engineering Society 131st International Conference*, 2007.

[25] ITU-R BS.2051-2, "Advanced Sound System for Programme Production," Standard, International Telecommunication Union, Geneva, CH, 2018.

[26] ITU-R BS.1116-3, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," Standard, International Telecommunication Union, Geneva, CH, 2015.