Remastering Divide and Remaster: A Cinematic Audio Source Separation Dataset with Multilingual Support

Karn N. Watcharasupat$^{*(\star),**}$, Chih-Wei Wu*, and Iroro Orife*,
*Audio Algorithms, Netflix, Inc., Los Gatos, CA 95032, USA (*Internship)
**Music Informatics Group, Georgia Institute of Technology, Atlanta, GA 30332, USA
Email: kwatcharasupat@gatech.edu, {chihweiw, iorife}@netflix.com

***Abstract*—Cinematic audio source separation (CASS), as a problem of extracting the dialogue, music, and effects stems from their mixture, is a relatively new subtask of audio source separation. To date, only one publicly available dataset exists for CASS, that is, the Divide and Remaster (DnR) dataset, which is currently at version 2. While DnR v2 has been an incredibly useful resource for CASS, several areas of improvement have been identified, particularly through its use in the 2023 Sound Demixing Challenge. In this work, we develop version 3 of the DnR dataset, addressing issues relating to vocal content in non-dialogue stems, loudness distributions, mastering process, and linguistic diversity. In particular, the dialogue stem of DnR v3 includes speech content from more than 30 languages from multiple families including but not limited to the Germanic, Romance, Indo-Aryan, Dravidian, Malayo-Polynesian, and Bantu families. Benchmark results using the Bandit model indicated that training on multilingual data yields significant generalizability to the model even in languages with low data availability. Even in languages with high data availability, the multilingual model often performs on par or better than dedicated models trained on monolingual CASS datasets.**

I. Introduction

Cinematic audio source separation (CASS), as a problem of extracting dialogue (DX), music (MX), and effects (FX) stems from their mixture, is a relatively young subtask of audio source separation. Unlike domain-specific source separation such as speech or music source separation, the audio content tackled by CASS spans essentially all possible natural and artificial sounds that could be recorded, synthesized, or otherwise produced. While CASS shares some similarities with universal audio source separation [1], CASS deals with audio source separation within an inherently creative domain, with a context-dependent and malleable categorization of sound classes into the three-stem setup. In particular, many musical instruments can be used to create non-musical sound effects. In this context, these sounds from musical instruments should be placed into the FX stem. Similarly, many objects typically not considered musical instruments have often been used in a musical context and their sounds should be placed into the MX stem, as per the contextual usage of the sounds. As expected with artistic uses of sounds, many additional edge cases and variations in practice exist when it comes to the categorization of sound into the DX, MX, and FX stems[1].

Efforts to improve the perceived sound quality or intelligibility of cinematic audio content date back to at least [2]. In the context of streaming media over the internet, CASS can act as a versatile preprocessor that opens up possibilities for audience-side personalization of cinematic content [3, 4]. The idea of CASS as a three-stem source separation problem was first introduced by Pétermann et al. [5, 6], together with the Divide and Remaster (DnR) dataset, a synthetic dataset with raw content drawn from LibriSpeech [7], FMA [8], and FSD50K [9]. DnR was then used in the Cinematic Audio Demixing (CDX) track of the 2023 Sound Demixing Challenge [10], with participation from various academic and industry groups. In the process, several areas of improvement to DnR, still the only publicly available CASS dataset, have been suggested. Specifically, Uhlich et al. [10] and other researchers have identified potential issues concerning the mismatch in production quality of the data from real cinematic audio, the lack of emotional content of the dialogue data, and vocal contents in non-dialogue stems. Our informal internal listening test of the Bandit model [11] on real Netflix content also indicated that the model, trained on DnR v2, often struggles with non-English dialogue, confusing unseen phonemes and use of linguistic tones as music or effects.

In this work, we developed version 3 of the DnR dataset[2]. While the dataset remains synthetic, it is our hope that version 3 now more closely reflects not just the common but also the diverse variations in cinematic audio production across languages, regions, genres, and creative practices. The major changes from DnR v2 are as follows:

- the dialogue stem now contains content from more than 30 languages across various language families;
- speech, vocals, and/or vocalizations have been removed from the music and effects stems;
- loudness and timing parametrization have been adjusted to approximate the distributions of real cinematic content; and,
- the mastering process now preserves relative loudness between stems and approximates standard industry practices.

As a benchmark, we trained Bandit models [11] on multiple linguistic variants of the proposed dataset and demonstrated that the multilingual model often performs on par or better

[1] See tinyurl.com/nflx-mne-guidelines for examples.

[2] Landing page: github.com/kwatcharasupat/source-separation-landing

than dedicated monolingual models on their respective test sets. We have made efforts to ensure that the proposed dataset was derived entirely from contents that permit commercial use, derivative work, and redistribution. The resulting dataset will be released under the CC BY-SA 4.0 License, while the replication code will be released under the Apache 2.0 license. Full attributions will be provided in the repository. The authors hope that the dataset will continue to be improved with more diverse sonic inventories for all three stems as more applicable datasets are developed and/or identified.

II. Dataset Setup

The Divider and Remaster dataset v3 consists of three splits and multiple linguistic variants each: train (6000 clips per variant), validation (600 clips per variant), and test (1200 clips per variant). Each clip is a collection of DX, MX, foreground FX (FGFX), background FX (BGFX), their mixture, and the combined FX stems. The dialogue stem in each variant draws from a different language or group of languages. For each identically indexed clip across variants, the music and effects stem share identical underlying pre-mastering tracks, but will differ slightly in the final tracks due to the mastering that depends on the mixture loudness.

All audio data are mono and $D_{\text{track}} = 60\,\text{s}$ in duration, as with DnR v2. However, audio data in v3 are sampled at 48 kHz with a bit depth of 24 bit, in line with delivery specifications used by major cinematic content providers such as Netflix, Sony Pictures, Warner Bros. Discovery, and Hulu[3].

A. On the Lack of Spatialization

Although most cinematic content is multichannel, DnR v3 remains a one-channel dataset. The rationale for this decision is due to the very open problem of *algorithmically* generating realistic source trajectories. Moreover, even though many methods for artificial spatialization exist for arbitrary trajectories and channel layouts, they are often unable to recreate realistic spatialized reverberations without relying on human-in-the-loop artistic intervention. Providing DnR v3 beyond mono without meaningful spatialization would provide little additional benefits in teaching a model to address spatial coherence issues. Therefore, we decided to remain in one channel for the time being. Multichannel inference can still be performed on models trained on mono data by treating each channel as a pseudo-mono track, albeit likely with some phase coherence issues.

B. License Consideration

Our best efforts have been made to check that the data used in this work permits the creation of derivative works, commercial use, or redistribution of the original/derivative works. For datasets where the licenses were specified at the track level, we have verified each license individually and excluded any file that does not meet the above requirements from further development in this work. Any data files with missing or unclear licenses were also excluded.

[3] See tinyurl.com/{nflx, sony, wbd, hulu}-delivery-specs.

TABLE I
Distribution of the licenses of the audio tracks in FMA.

Com.?	Deriv.?	License	Count	%
Y	Y	CC BY (1.0, 2.0, 2.5, 3.0, 4.0)	6960	6.5
		CC BY-SA (2.0, 2.5, 3.0, 4.0)	2802	2.6
		Public Domain	1392	1.3
		Free Art License	16	0.02
		Open Audio License	5	0.005
	N	CC BY-ND (2.0, 2.5, 3.0, 4.0)	962	0.9
	?	CC Sampling+ (1.0)	14	0.01
N	Y	CC BY-NC (2.0, 2.1, 2.5, 3.0, 4.0)	8313	7.8
		CC BY-NC-SA (2.0, 2.1, 2.5, 3.0, 4.0)	43 512	40.8
	N	CC BY-NC-ND (2.0, 2.1, 2.5, 3.0, 4.0)	41 869	39.3
		Free Music Philosophy	138	0.1
		CopyrightPlus	23	0.02
		FMA Music Sharing	18	0.02
		ideology.de License	10	0.009
	?	CC NC-Sampling+ (1.0)	19	0.02
?	?	Sound Recording Common Law	428	0.4
		Missing license	88	0.08
		Orphan work	5	0.005

Com.?: Allows commercial use?; Deriv.?: Allows derivative works?

C. Music Data and Preprocessing

As with DnR v2, we utilize the Free Music Archive (FMA) dataset [8] for the music stem. FMA provided MP3-encoded stereo audio data sampled at 44.1 kHz. The audio tracks in FMA are individually licensed thus only a subset permits commercial and derivative use. Specifically, only public-domain audio tracks and those licensed under the CC BY, CC BY-SA, Open Audio License, or Free Art License families were included. The distribution of the licenses of FMA audio tracks can be found Tab. I. In total, 11 175 (10.5 %) out of 106 574 tracks were included for further development.

Due to the significant reduction in the number of available tracks, we utilized the full-track version of FMA instead of the 30-second version used in DnR v2. Segments with speech and/or vocals were removed by running a speech-music activity detection (SMAD) model [12] on the track and keeping only contiguous segments without any speech activation with a minimum duration of 5.0 s. To maximize data diversity in terms of mixing parameters, we treated the left and right channels of the data as two pseudo-independent mono tracks, instead of downmixing them into mono tracks as done in DnR v2. After the preprocessing, we have 170 386 segments of music data without vocal or speech content, totaling 1143.5 h.

D. Effects Data and Preprocessing

As with DnR v2, we also utilized the *FSD50K* dataset [9]. Audio tracks in FSD50K are also individually licensed with either CC0, CC BY, CC BY-NC, or CC Sampling+. Only the CC0 and CC BY licensed tracks were used in v3. After filtering, 43 379 (84.7 %) out of 51 197 tracks were included.

In order to filter out tracks that may include speech or music content, we first refer to the AudioSet ontology [13] used in FSD50K. Any track under the "Human Voice", "Music", and "Sound reproduction" umbrellas were excluded. Within the "Human group actions" umbrella, all tracks except those explicitly labeled as “Clapping” or “Applause” were excluded. Exclusion criteria take priority in tracks with multiple labels.

Of the 25 147 remaining tracks, we further ran SMAD on any file longer than 1 s in duration and excluded those with any detected speech or music frame. After the preprocessing, we have 21 760 segments of non-vocal FX data, totaling 41.3 h.

Unlike DnR v2, however, any sound from any class can be either foreground or background. The only differences between the FGFX and BGFX stems lie in the loudness and density of sound events.

E. Dialogue Data and Preprocessing

The dialogue stem in DnR v3 consists of data from 32 languages across 40 datasets, summarized in Tab. II. The datasets were chosen by first going through all available datasets in OpenSLR that satisfy the quality and license requirements[4]. We then perform naive web searches for any dataset representing a language within the 200 most spoken languages, as per the 2023 Ethnologue 200, that satisfies the requirements.

Given the large variety of data formats, any included audio track is required to have a sampling rate of at least 44.1 kHz, a bit depth of at least 16 bits for lossless encoding, or a bit rate of at least 64 kbps for lossy encoding. Manual spot checks were performed on each dataset. Datasets with existing sentence- or utterance-level segmentation were kept as is. Longer audio files were segmented, with leading and trailing silences removed for each segment. Datasets with existing train-validation-test or equivalent splits were respected regardless of the original distributions. Datasets with only train-test splits or equivalent were split into train-validation-test splits by breaking up the original train split into a train split and validation split at speaker level. Any dataset with only one speaker was treated as a train-only dataset. Any dataset with between two to six speakers was split, if possible, to have the test set consist of at least one male and one female speaker, followed by the train and then validation sets, in this order of priority. Dataset-specific details along with linguistic background are detailed below. Linguistic and demographic information was drawn from Ethnologue [34] unless explicitly cited. All included Google datasets (SLR 32, 37, 41–44, 61, 63–66, 69–80, 83, 86) are licensed under CC BY-SA 4.0.

English (ENG) is a West Germanic language with more than 1.5 B total speakers across the world, making it the most spoken language in the world in terms of total speakers[5]. Due to its widespread use throughout the world, significant regional varieties of English have emerged, including many dialects, creoles, and pidgins, with significant variations in phonology. In general, English is a non-tonal and stress-timed language, with 24 consonants, 13 vowels, and 8 diphthongs. English data were drawn from LibriSpeech (SLR 12), SLR 70, and SLR 83. LibriSpeech [7] is a large-scale dataset of read English audiobook speech drawn from the LibriVox project. As with DnR v2, the original 44.1 kHz MP3 files from LibriVox were used for the creation of DnR v3. LibriSpeech is licensed under CC BY 4.0 while the original LibriVox data is public domain. SLR 70 is a dataset of Nigerian English, without pidgin, recorded in Lagos, Nigeria, and London, UK [14]. Based on British English, Nigerian English contains prosodic features closer to a tonal language [35]. Note, however, that Nigerian English consists of multiple regional varieties with differing distinctive features. SLR 83 is a dataset of various English dialects across the British Isles, namely Irish, Scottish, Welsh, Midlands, Northern England, and Southern England.

Standard German (DEU) is also a West Germanic language with more than 130 M total speakers, mainly in Western and Central Europe. German is a non-tonal and stress-timed language, with 22 consonants, 22 vowels, and 3 diphthongs, notably with the rare voiceless labiodental affricate /pf/. German data were drawn from the HUI Audio Corpus [16] which was also derived from LibriVox. The dataset, with almost 120 speakers, is licensed under CC0. Although unstated in [16], these audio were likely originally obtained from LibriVox as lossily encoded MP3 files with a bit rate of 128 kbps.

Faroese (FAO) is a North Germanic language with 69 K total speakers, mostly in the Faroe Islands. Descended from Old Norse (NON), Faroese has around 20 consonants, more than 20 vowels in some analyses, and 8 diphthongs [36]. The dataset for Faroese is the Basic Language Resource Kit (BLARK) for Faroese (SLR 125) [17], licensed under CC BY 4.0. Faroese data were provided unsegmented. Pydub was used to segment the Faroese data into nonsilent segments.

French (FRA) is a Romance language, specifically from the Gallo-Romance branch, with more than 300 M total speakers across all continents. As with all Romance languages, French descended from Vulgar Latin. It is a non-tonal language with syllable-timed stress, consisting of 20 consonants and 14 vowels. French features guttural R, nasal vowels, and liaison between words. French data were drawn from the AudioCité corpus (SLR 139) [18], an extremely large read speech corpus with more than 6700 hours of data. AudioCité data are individually licensed and provided in an MP3 format at various bit rates and sampling rates. As a result, we only included sufficiently high-quality data that are public domain, or licensed under CC BY, CC BY-SA, or License Art Libre. AudioCité data were provided unsegmented, with some sections containing non-vocal content such as introductory music. SMAD was used to detect non-silent speech segments of at least 1 s in duration and to remove any segment with music. After preprocessing, 1.3 M segments from 58 unique speakers, totaling 3478 h, remain for further development.

Italian (ITA) is a Romance language from the Italo-Dalmatian branch, with 67 M total speakers, primarily in Italy. It is a non-tonal language with 23 consonants and 7 vowels. Compared to other Romance languages, Italian preserves much of the Vulgar Latin phonology. Italian is represented by the LaMIT Database [19], licensed under CC BY 4.0.

Catalan (CAT) is a Romance language from the Ibero-Romance branch, with 9.3 M total speakers, primarily in Spain.

[4] openslr.org. The latest entry at the time of writing was SLR 151.

[5] “Total speakers” refers to the total number of “first-language” (L1) and “second-language” (L2) speakers.

TABLE II
LIST OF LANGUAGES INCLUDED IN DNR V3

					Original Format			Voices			
Language		Family		Source Dataset	f_s (kHz)	Codec	Hrs.	F	M	U	L1+L2 (M)
English, unspecified	ENG	W. Germanic	GMW	LibriSpeech [7]	44.1	MP3, ~128 kbps	111.4	165	166		1515
English, Nigerian	ENG	W. Germanic	GMW	SLR 70 [14]	48	PCM, 16-bit	5.8	19	12		
English, Irish	ENG	W. Germanic	GMW	SLR 83 [15]	48	PCM, 16-bit	0.7		3		
English, Scottish	ENG	W. Germanic	GMW	SLR 83 [15]	48	PCM, 16-bit	4.3	6	11		
English, Welsh	ENG	W. Germanic	GMW	SLR 83 [15]	48	PCM, 16-bit	5.4	8	11		
English, Midlands	ENG	W. Germanic	GMW	SLR 83 [15]	48	PCM, 16-bit	1.2	2	3		
English, N. England	ENG	W. Germanic	GMW	SLR 83 [15]	48	PCM, 16-bit	5.0	5	14		
English, S. England	ENG	W. Germanic	GMW	SLR 83 [15]	48	PCM, 16-bit	14.6	28	29		
German, Standard	DEU	W. Germanic	GMW	HUI [16]	44.1	PCM, 16-bit	253.8	61	54	3	134
Faroese	FAO	N. Germanic	GMQ	SLR 125 [17]	48	PCM, 16-bit	94.5			369	<1
French	FRA	Romance	ROA	Audiocite [18]	44.1, 48	MP3, 64–320 kbps	3478.3	32	22	4	312
Italian	ITA	Romance	ROA	LaMIT [19]	44.1	PCM, 16-bit	0.8	2	2		67
Catalan	CAT	Romance	ROA	SLR 69 [20]	48	PCM, 16-bit	9.4	20	16		9
Galician	GLG	Romance	ROA	SLR 77 [20]	48	PCM, 16-bit	10.3	34	10		3
Spanish, Argentinian	SPA	Romance	ROA	SLR 61 [21]	48	PCM, 16-bit	8.1	31	13		560
Spanish, Peninsular	SPA	Romance	ROA	SLR 61 [21]	48	PCM, 16-bit	0.1	1	2		
Spanish, Chilean	SPA	Romance	ROA	SLR 71 [21]	48	PCM, 16-bit	7.1	13	18		
Spanish, Colombian	SPA	Romance	ROA	SLR 72 [21]	48	PCM, 16-bit	7.6	16	17		
Spanish, Peruvian	SPA	Romance	ROA	SLR 73 [21]	48	PCM, 16-bit	9.2	18	20		
Spanish, Puerto Rican	SPA	Romance	ROA	SLR 74 [21]	48	PCM, 16-bit	1.0	5			
Spanish, Venezuelan	SPA	Romance	ROA	SLR 75 [21]	48	PCM, 16-bit	4.8	11	12		
Basque	EUS	Paleo-European		SLR 76 [20]	48	PCM, 16-bit	13.9	29	23		1
Kazakh	KAZ	Turkic	TRK	SLR 140 [22]	44.1, 48	PCM, 16-bit	11.5			22	17
Ukranian	UKR	E. Slavic	ZLE	[23]	48	PCM, 16-bit	16.6	2	1		39
						OPUS, ~128 kbps	6.3		1		
Arabic, Levantine	APC	Semitic	SEM	ASC [24]	48	PCM, 16-bit	4.1		1		54
Bengali	BEN	Indo-Aryan	INC	SLR 37 [25]	48	PCM, 16-bit	5.0		15		278
Nepali	NPI	Indo-Aryan	INC	SLR 43 [25]	48	PCM, 16-bit	2.8	18			32
Marathi	MAR	Indo-Aryan	INC	SLR 64 [26]	48	PCM, 16-bit	3.0	9			99
Gujarati	GUJ	Indo-Aryan	INC	SLR 78 [26]	48	PCM, 16-bit	7.9	18	18		63
Malayalam	MAL	Dravidian	DRA	SLR 63 [26]	48	PCM, 16-bit	5.5	24	18		38
Tamil	TAM	Dravidian	DRA	SLR 65 [26]	48	PCM, 16-bit	7.1	25	25		87
Telugu	TEL	Dravidian	DRA	SLR 66 [26]	48	PCM, 16-bit	5.7	24	23		96
Kannada	KAN	Dravidian	DRA	SLR 79 [26]	48	PCM, 16-bit	8.5	23	36		59
Malay	ZLM	Malayo-Polynesian	POZ	Malaya Speech [27]	44.1	PCM, 16/24-bit	6.7	1	1		19
Javanese	JAV	Malayo-Polynesian	POZ	SLR 41 [25]	48	PCM, 16-bit	7.0	19	20		68
Sundanese	SUN	Malayo-Polynesian	POZ	SLR 44 [25]	48	PCM, 16-bit	5.4	20	21		32
Khmer	KHM	Mon-Khmer	MKH	SLR 42 [25]	48	PCM, 16-bit	4.0		16		18
Burmese	MYA	Tibeto-Burman	TBQ	SLR 80 [28]	48	PCM, 16-bit	4.1	20			43
Chinese, Mandarin	CMN	Chinese	ZHX	AISHELL-3 [29]	44.1	PCM, 16-bit	85.6	175	43		1140
Japanese	JPN	Japonic	JPX	JVNV [30]	48	PCM, 24-bit	3.9	2	2		123
				PJS [31]	48	PCM, 24-bit	0.6		1		
Afrikaans	AFR	W. Germanic	GMW	SLR 32 [32]	48	PCM, 16-bit	3.3			9	18
Xhosa	XHO	Bantu	BNT	SLR 32 [32]	48	PCM, 16-bit	3.1			6	19
Sesotho	SOT	Bantu	BNT	SLR 32 [32]	48	PCM, 16-bit	3.2			6	14
Setswana	TSN	Bantu	BNT	SLR 32 [32]	48	PCM, 16-bit	3.5			5	14
Yoruba	YOR	Defoid		SLR 86 [33]	48	PCM, 16-bit	4.0	19	17		47

Languages are grouped by genealogical proximity and/or geopolitical proximity of their primary region(s) of use or origin.

It is a non-tonal language with 22 consonants, 7 vowels, and 4 diphthongs, inheriting the vowel system from Vulgar Latin. Represented by SLR 69 [20], Catalan features terminal devoicing, lenition, voicing assimilation, and vowel harmony.

Galician (GLG) is also a Romance language in the Ibero-Romance branch, with 3.4 M total speakers, primarily in Spain. Galician shares an ancestor with Portuguese (POR), both originating from the Galician-Portuguese language, thus their significant mutual intelligibility. Galician features 21 consonants, 7 vowels, and 9 diphthongs [37], also inheriting the vowel system from Vulgar Latin. Like Catalan and some Portuguese dialects, Galician features vowel harmony. Galician is represented by SLR 67 [20].

Spanish (SPA) is a Romance language from the Ibero-Romance branch, with 559 M total speakers throughout the world, especially in Latin America, making it the second-most spoken language in terms of native speakers. Spanish is a non-tonal and syllable-timed language with 20 consonants, 5 vowels, and 5 diphthongs. Spanish phonology varies significantly across dialects and regions. Spanish data were drawn from six datasets recorded in Latin America [21], specifically, Argentina (SLR 61), Chile (SLR 71), Colombia (SLR 72), Peru (SLR 73), Puerto Rico (SLR 74), and Venezuela (SLR 75). SLR 61 includes a small amount of weather announcements in Peninsular Spanish.

Basque (EUS) is a language isolate with 1 M total speakers, primarily in Spain. The only surviving member of the Paleo-European languages, Basque is a non-tonal language with 24 consonants, 5 vowels, and 6 diphthongs. It is represented by SLR 66 [20].

Kazakh (KAZ) is Western Turkic language with 16 M total speakers. Kazakh is a non-tonal language, with 18 consonants and 9 vowels. As with many Turkic languages, Kazakh exhibits vowel harmony. It is represented by the Kazakh Speech Dataset (SLR 140) [22], licensed under CC BY-SA 3.0.

Ukrainian (UKR) is an East Slavic language with 39 M total speakers, primarily in Ukraine. It is a non-tonal language, with 32 consonants and 6 vowels. The data for Ukrainian is drawn from the Lada, Tetiana, Mykyta, and Oleksa text-to-speech (TTS) datasets in [23], all licensed under Apache 2.0.

Levantine Arabic (APC) is a variety of Arabic (ARA spoken in the Levant region, with 54 M total speakers. While it shares similarities with the 28 Modern Standard Arabic (ARB) consonants, vowels differ by dialects, usually with 5 short vowels, 5 long vowels, and 2 diphthongs [38]. Data were drawn from the Arabic Speech Corpus [24], licensed under CC BY-SA 4.0, recorded by a single male speaker.

Bengali (BEN) is an Eastern Indo-Aryan language from the Bengali-Assamese branch, with 278 M total speakers, primarily in the Bengal region of the Indian subcontinent. As with most Indo-Aryan languages, it is non-tonal. It has 35 consonant [34], 7 vowels all with nasalization [39], and many diphthongs [40]. Bengali data were drawn from SLR 37 [25].

Nepali (NPI) belongs to the Eastern Pahari branch, with 32 M total speakers, primarily in Nepal. It is a non-tonal language, with 29 consonants, 6 oral vowels, 5 nasal vowels, and 10 diphthongs. Nepali data were drawn from SLR 43 [25].

Marathi (MAR) is a Southern Indo-Aryan language with 99 M total speakers, primarily in the Maharashtra state in Western India. It is a non-tonal language, with more than 30 consonants and 12 vowels, including 2 diphthongs [41]. Marathi does not feature nasal vowels. Marathi data were drawn from SLR 64 [26].

Gujarati (GUJ) is a Western Indo-Aryan language with 63 M total speakers, primarily in the Gujarat state in Western India. It is a non-tonal language with 31 consonants, 8 vowels, and 2 diphthongs. Gujarati data were drawn from SLR 78 [26].

Malayalam (MAL) is a Southern Dravidian language with 38 M total speakers, primarily in the Kerala state on the Southwestern coast of India. It is a non-tonal language, with 37 consonants, 11 vowels, and 2 diphthongs. As with other Dravidian languages, Malayalam features true subapical retroflexes. Malayalam is represented by SLR 63 [26].

Tamil (TAM) is a Southern Dravidian language with 87 M total speakers, primarily in the Tamil Nadu state in Southern India. It is a non-tonal language with 18 consonants, 10 vowels, and 2 diphthongs. In addition to the true subapical retroflexes, Tamil features multiple rhotic consonants. Tamil is represented by SLR 65 [26].

Telugu (TEL) is a South-Central Dravidian language with 96 M total speakers, primarily in the Andhra Pradesh and Telangana states in Southern India. A non-tonal language with 21 consonants and 11 vowels, it is represented by SLR 66 [26].

Kannada (KAN) is a Southern Dravidian language with 59 M total speakers, primarily in the Karnataka state in Southwestern India. It is a non-tonal language with, 22 consonants, 20 vowels, and 2 diphthongs. It is represented by SLR 79 [26].

Mandarin Chinese (CMN) is a Sino-Tibetan language with more than 1.1 B total speakers, primarily in China. Mandarin Chinese is the most commonly spoken language in terms of native speakers. Chinese is a tonal language with 4 phonemic tones, 24 consonants, 8 vowels, and 6 diphthongs. The dataset for Mandarin Chinese is AISHELL-3 (SLR 93) [29], licensed under Apache 2.0.

Japanese (JPN) is a Japonic language with 123 M total speakers, primarily in Japan. It is a non-tonal but pitch-accented language with 15 consonants, 5 vowels, and 3 diphthongs. Data for Japanese were drawn from emotional speech JVNV corpus [30], and the read speech subset of PJS Corpus [31]. Both datasets are licensed under CC BY-SA 4.0.

Burmese (MYA) is a Sino-Tibetan language from the Tibeto-Burman branch with 43 M total speakers, primarily in Myanmar. It is a tonal language with 3 primary tones, 31 consonants, 8 vowels, and 4 diphthongs. It is represented by SLR 80 [28].

Khmer (KHM) is an Austro-Asiatic language from the Mon-Khmer family with 18 M total speakers, primarily in Cambodia. Unlike most continental Southeast Asian languages, it is a non-tonal language. Khmer has 44 consonants and 24 vowels [42]. Khmer is represented by SLR 42 [25].

Malay (ZLM) is an Austronesian language from the Malayo-Polynesian branch with 19 M total speakers, primarily in Malaysia. It is part of the Malay macrolanguage (MSA) used throughout maritime Southeast Asia that includes Standard Malay (ZSM) and Indonesian (IND). Like most Austronesian languages, Malay is non-tonal, with 19 consonants and 6 vowels [43]. Data for Malay were drawn from datasets released via Malaya Speech [27], predominantly recorded by one male speaker, all licensed under CC BY 4.0.

Javanese (JAV) is an Austronesian language from the Malayo-Polynesian branch with 68 M total speakers, primarily in the Java and Sumatra islands of Indonesia. Javanese is a non-tonal language with 21 consonants and 8 vowels. It is represented by SLR 41 [25].

Sundanese (SUN) is an Austronesian language from the Malayo-Polynesian branch with SI37M total speakers, primarily in the Western part of the Java island in Indonesia. Sundanese is a non-tonal language with 18 consonants and 7 vowels. Sundanese is represented by SLR 44 [25].

Afrikaans (AFR) is a West Germanic language with 18M speakers, primarily in South Africa. It is highly related to Dutch (NLD) due to the Dutch colonization of South Africa. It is a non-tonal language with 20 consonants, 16 vowels, and 9 diphthongs. Afrikaans is represented by SLR 32 [32].

Xhosa (XHO) is a Bantu language from the Nguni branch, with 19 M total speakers, primarily in South Africa. It is a tonal language with 2 phonemic tones (high and low) and 10 vowels. Xhosa notably has a complex system of consonants, with multiple ejective stops and click consonants [44]. It is also represented by SLR 32 [32].

Sesotho (SOT) is a Bantu language from the Sotho-Tswana branch, with 13.5 M total speakers, primarily in South Africa and Lesotho. It is a tonal language with 2 tones, more than 30 consonants including 3 clicks, and 9 vowels [45]. It is also represented by SLR 32 [32].

Setswana (TSN) also belongs to the Sotho-Tswana branch, with 14 M user, primarily in South Africa and Botswana. It is a tonal language with two tones, 29 consonants, and 7 vowels [46]. It is also represented by SLR 32 [32].

Yoruba (YOR) is an Edekiri language with 47 M total speakers, primarily in Nigeria. A tonal language with 3 tones, 17 consonants, 7 oral vowels, and 4 nasal vowels, Yoruba features vowel harmony. It is represented by SLR 86 [33].

III. Dataset Generation

The process for generating the stem track is broadly similar to that of DnR v2. The precise processes are detailed in Algo. 1 and Algo. 2. The process at the high level is as follows.

For each stem, the track-level loudness L_{track} and number of sound events N_{event} were randomly drawn. For each sound event, up to 10 event candidates were randomly drawn from the pool of raw sound events and checked against timing constraints. The first raw event that satisfies the timing constraints is taken. The timing parameters (start time within track t_i, start time within event η_i, and event duration U_i) were then set or randomized accordingly. If none of the candidates satisfy the constraints, that event is skipped. The event-level loudness $L_{\text{event},i}$ was then drawn from a distribution centered at the track-level loudness. The raw sound event was segmented and loudness-normalized accordingly, before being added to the stem. Once all events had been added, the stem was loudness normalized again to the track-level loudness. Finally, mastering is applied to meet target loudness and peak specifications as detailed in Section III-E. All mentions of "loudness" in this work refer to the integrated loudness as defined in ITU-R BS.1770 [47].

We will now discuss the details and motivation behind each step and any changes from version 2. Stem-dependent parameters are provided in Tab. III.

A. Loudness distribution

In DnR v3, the track-level loudness L_{track} is drawn from $\mathcal{N}(\mu_{\text{ref}}+\Delta\mu_{\text{track}}, \sigma^2_{\text{track}})$, where $\mu_{\text{ref}} = -27$ LKFS is an arbitrary reference level, $\Delta\mu_{\text{track}}$ is a stem-dependent mean loudness level, and σ_{track} is a stem-dependent track-level loudness standard deviation (SD). Unlike DnR v2, μ_{ref} does not affect peak limiting during mastering. The event-level loudness L_{event} was drawn from $\mathcal{N}(L_{\text{track}}, \sigma^2_{\text{event}})$, where σ_{event} is a stem-dependent event-level loudness SD. Note that once the entire stem is generated, the entire track is normalized back to L_{track}.

For simplicity, we set $\Delta\mu_{\text{track}} = 0$ LKFS for the DX stem, $\Delta\mu_{\text{track}} = -13.0$ LKFS for the BGFX stem, and $\Delta\mu_{\text{track}} = -5$ LKFS for the music and FGFX stem. The level for music and foreground stems relative to the DX stem are approximately based on the observed distributions both in Netflix's own content distribution and in CDXDB23 [10, Tab. 6] which was derived from Sony content. Setting the BGFX stem at 8 LKFS under the FGFX is based on DnR v2. Event loudness normalization is performed using pyloudnorm [48], which implements ITU-R BS.1770-4 [47].

TABLE III
Stem-dependent parameters

Parameter	DX	MX	FGFX	BGFX
λ_{event}	12.0	7.0	12.0	24.0
$\Delta\mu_{\text{track}}$ (LKFS)	0.0	−5.0	−5.0	−13.0
σ_{track} (LKFS)	4.0	6.0	6.0	6.0
σ_{event} (LKFS)	6.0	10.0	10.0	10.0
T_{min} (s)	0.0	0.0	0.5	1.0
β_{min}	1.0	0.3	0.3	0.3
γ	0.75	1.0	0.5	0.0

Algorithm 1: Stem Generation

```
Initialize buffer $\mathbf{s}$ to zeros of duration $D_{\text{track}}$.
Initialize track cursor $t_{\text{cur}} \leftarrow 0\,\text{s}$.
Sample track loudness $L_{\text{track}} \sim \mathcal{N}(\mu_{\text{ref.}} + \Delta\mu_{\text{track}}, \sigma^2_{\text{track}})$.
Sample number of events $N_{\text{event}} \sim \mathcal{ZTP}(\lambda_{\text{event}})$.
for $i := 1$ to $N_{\text{event}}$ do
    for up to 10 trials do
        Sample raw sound event $\mathbf{z}_i \sim \mathfrak{Z}$.
        Run Algo. 2 on $\mathbf{z}_i$ to check and get timing parameters.
        if Algo. 2 succeeds then
            Set $t_i$, $\eta_i$, and $U_i$ accordingly.
            Sample event loudness $L_{\text{event},i} \sim \mathcal{N}(L_{\text{track}}, \sigma^2_{\text{event}})$.
            Segment $\tilde{\mathbf{z}}_i \leftarrow \mathbf{z}_i(\eta_i : \eta_i + U_i)$
            Normalize loudness of $\tilde{\mathbf{z}}_i$ to $L_{\text{event},i}$.
            Set $\mathbf{s}(t_i : t_i + U_i) \leftarrow \mathbf{s}(t_i : t_i + U_i) + \tilde{\mathbf{z}}_i$.
            Sample $\Delta t_{\text{cur}} \sim \mathcal{U}(\gamma U_i, U_i)$.
            Set $t_{\text{cur}} \leftarrow t_{\text{cur}} + \Delta t_{\text{cur}}$.
            break
Normalize loudness of $\mathbf{s}$ to $L_{\text{track}}$.
```

Two major changes from version 2 are in the significantly increased spread of the loudness distribution and the underlying distribution itself. In DnR v2, L_{track} was drawn from a uniform distribution with bounds ±2 LKFS from the mean, while L_{event} was drawn uniformly with bounds ±1 LKFS from L_{track}. The resulting triangle distribution can be seen to be much narrower than the distribution of the loudness in CDXDB23 [10, Fig. 6]. To better approximate the loudness distribution, we switched to the normal distribution in version 3. The fairly large event-level SDs were set so as to also help mimic the effects of panning on the loudness, hopefully allowing the models trained on it to better function in a pseudo-stereo mode.

B. Event density

As with DnR v2, the number of sound events $N-$ event within each track is drawn from a zero-truncated Poisson distribution $\mathcal{ZTP}(\lambda_{\text{event}})$, where λ_{event} is a stem-dependent parameter for the mean number of events. The values of λ_{event} for MX and FGFX are the same as those of DnR v2. The value of λ_{event} for BGFX is set very high, so that, in tandem with the low relative loudness, the stem provides a dense but unsalient backdrop to the mixture. The value of λ_{event} for DX is set slightly higher than that of DnR v2 to account for the

Algorithm 2: Check and get timing parameters

```
Input: Raw event $\mathbf{z}_i$, Cursor $t_{\text{cur}}$.
Set $T_i \leftarrow$ duration of $\mathbf{z}_i$, in seconds.
Set $S_{\text{min}} \leftarrow \max\{T_{\text{min}}, \beta T_i\}$.
Set $t_{\text{max}} \leftarrow D_{\text{track}} - S_{\text{min}}$.
if $D_{\text{track}} - t_{\text{cur}} < S_{\text{min}}$ then return FAIL.
if $\beta_{\text{min}} < 1$ then
    Clamp $t_{\text{max}} \leftarrow \min\{t_{\text{max}}, D_{\text{track}} - \delta\}$.
    if $t_{\text{max}} < t_{\text{cur}}$ then return FAIL.
Sample start time in track $t_i \sim \mathcal{SN}(t_{\text{cur}}, \rho^2_{\text{start}}, \alpha)$.
if $\beta_{\text{min}} = 1$ or $T_i \leq S_{\text{min}}$ then
    Clamp $t_i$ to range $[t_{\text{cur}}, t_{\text{max}}]$.
    Set event duration $U_i \leftarrow T_i$.
else
    Set $S_{\text{max}} \leftarrow \min\{T_i, D_{\text{track}} - t_i\}$
    if $S_{\text{max}} < 0$ then return FAIL.
    Clamp $t_i \leftarrow \max\{t_i, 0\,\text{s}\}$.
    Sample event duration $U_i \sim \mathcal{TN}(\upsilon T_i, (\rho T_i)^2, S_{\text{min}}, S_{\text{max}})$.
if MX then
    Sample event start time offset $\eta_i \sim \mathcal{U}(0\,\text{s}, T_i - U_i)$.
else
    Set event start time offset $\eta_i = 0\,\text{s}$.
return $t_i$, $\eta_i$, $U_i$
```

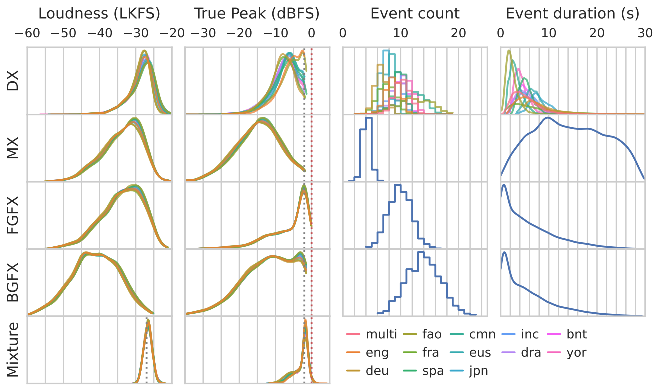

Fig. 1. Test set distribution of the post-mastering loudness, true peak, event counts, and event durations for each stem. Each colored line represents a variant. The dotted line in the mixture loudness plot indicates the −27 LKFS level. The dotted lines in the peak plots indicate −2 dBFS and 0 dBFS levels.

relatively shorter average length of raw sound events in some of the contributing dialogue datasets. In practice, it can be observed that the values of λ_{event} are higher than the actual mean number of events that can fit into each track given the timing constraints. Nonetheless, the distributions of the number of events generally follow Poisson-like shapes.

C. Sound event sampling

For each event, a random raw sound event candidate $\mathbf{z}_i$ is drawn randomly from the pool of possible sound events $\mathfrak{Z}$, where i is the event index. Sampling for dialogue events in monolingual variants, all music events, and all effects events were done uniformly with replacement. For the Indo-Aryan (INC), Dravidian (DRA), and Bantu (BNT) variants, a language is drawn with a probability defined by the ratio of the total number of speakers of the language relative to the total number of speakers of all represented languages in the respective variant. For the MULTI variant, languages from the West Germanic and Romance families were drawn according to the total number of speakers of each language. All other languages were drawn with a probability proportional to the total number of speakers of the language family. Within each language, each sample is drawn with an equal probability.

D. Timing Parameters

For each sound event candidate, the process detailed in Algo. 2 is applied to check whether the event can be placed into the stem, and if so, to compute the timing parameters, namely, the start time relative to the track t_i, the start time relative to the raw sound event η_i, and the event duration U_i. This process is very similar to that of DnR v2, except that a non-music sound event was allowed to overlap with the previous event by a duration of up to $(1-\gamma)U_{i-1}$, where γ is stem-dependent. At the start of the process, a track cursor t_{cur} is set to 0 s. After each event is added, t_{cur} is advanced by a random amount drawn from $\mathcal{U}(\gamma U_i, U_i)$.

The first constraint is that the remaining duration in a track is at least $U_{\text{min}} = \max\{T_{\text{min}}, \beta_{\text{min}} T_i\}$, where T_{min} is the absolute minimum duration, β_{min} is the minimum proportional duration, and T_i is the duration of $\mathbf{z}_i$. For the DX stem where $\beta_{\text{min}} = 1$, this ensures that the entire raw sound event must fit in the track to prevent a word from being cut off mid-enunciation. For the rest of the stems where $\beta_{\text{min}} < 1$, the latest possible start time $t_{\text{max}} = \min\{D_{\text{track}} - U_{\text{min}}, D_{\text{track}} - \delta\}$ must be later than t_{cur}, which approximately defines the earliest possible start time. The parameter $\delta = 2.0\,\text{s}$ follows from v2, and defines the minimum time before the end of the track that a non-dialogue event could start.

Once the first two constraints are satisfied. Similar to DnR v2, The start time relative to the track t_i is sampled from a skew-normal distribution $\mathcal{SN}(t_{\text{cur}}, \rho^2_{\text{start}}, \alpha)$ where $\rho_{\text{start}} = 2.0\,\text{s}$ defines the spread of the distribution and $\alpha = 5.0$ defines the skew parameter of the distribution. Since the skew-normal distribution is unbounded, t_i can take a value less than t_{cur} or larger than D_{track}, although with low probabilities. For the DX stem or events where $T_{\text{min}} < U_{\text{min}}$, t_i is clamped to the range $[t_{\text{cur}}, t_{\text{max}}]$ and the entire raw sound event is added to the stem, i.e. $U_i = T_i$. For other cases, if t_i is larger than D_{track}, we simply draw a new candidate. For the cases where $t_i < D_{\text{track}}$, we clamp t_i to ensure $t_i \geq 0\,\text{s}$. Following DnR v2, the event duration U_i is then drawn from a truncated normal distribution $\mathcal{TN}(\upsilon_{\text{dur}} T_i, (\rho_{\text{dur}} T_i)^2, U_{\text{min}}, U_{\text{max}})$ where $\upsilon_{\text{dur}} T_i$ is the location parameter, $\rho_{\text{dur}} T_i$ is the spread parameter, U_{min} is the lower bound, and U_{max} is the upper bound, with $\upsilon_{\text{dur}} = 0.5$ and $\rho_{\text{dur}} = 0.1$. For the MX stem with $U_i < T_i$, the start time offset η_i is randomized from a uniform distribution $\mathcal{U}(0\,\text{s}, T_i - U_i)$. All other stems start from the beginning of the raw sound event, i.e. $\eta_i = 0\,\text{s}$.

E. Mastering

Following Netflix's specifications of mastering to (−27 ± 1) LKFS, we draw the target loudness level from $\mathcal{N}(\mu_{\text{mix}}, \sigma^2_{\text{mix}})$ with $\mu_{\text{mix}} = -27\,\text{LKFS}$ and $\sigma_{\text{mix}} = 1\,\text{LKFS}$.

The choice of drawing the loudness randomly instead of targeting −27 LKFS directly is to reflect the occasional non-compliance with the delivery specifications seen in practice. We follow the true peak limit specification for Netflix and Hulu, which is −2 dBFS.

In practice, loudness normalization and peak limiting are usually applied on the mix bus. However, these processes are nonlinear and often time-variant. In order to keep the mixture as a linear sum of the stems, we apply loudness normalization and peak limiting to each stem individually.

In DnR v2, peak limiting was done using naive (sample) peak computation followed by attenuation of the entire stem. This, however, is not a common practice and can severely distort the relative loudness distribution between stems. In DnR v3, we first compute an intermediate mixture based on the generated stems. Using this intermediate mixture, we compute the necessary gain adjustment needed to get the mixture to the target loudness. Using this value, each stem is then linearly scaled to meet their respective absolute target loudness using pyloudnorm. We then use the loudnorm filter of FFmpeg to compute the necessary filter parameters in the first pass, and to jointly apply EBU R 128 loudness normalization and peak limiting [49] in the second pass, obtaining the final stems. Dynamic normalization is allowed if the first pass of loudnorm indicates that linear time-invariant normalization cannot achieve the required specifications. The final mixture is then computed by linearly combining the stems. Within any split of any variant, at most 2.7 % and 2.0 % of the mixtures are clipped in terms of true peak and sample peak, respectively.

IV. Experimental Setup

A. Model

The architecture of the model used in this work is identical to the 64-band music variant of Bandit in [11]. Note that the model in [11] was trained on data sampled at 44.1 kHz while the models in this work were trained on data sampled at 48 kHz. As a result, while the band definition in terms of the discrete frequency domain indices does not change, the actual band edges in terms of the absolute frequency in Hz are different.

B. Training and Testing

All models in this work were trained using the following setup, which is slightly different from [11]. Each model was trained for 200 epochs with a batch size of 8 per GPU, using an AdamW optimizer [50] with an initial learning rate of 10^{-3} and a decay factor of 0.99 per epoch. Each epoch consisted of 2048 batches per GPU. For each training clip, a random chunk of 8.0 s was drawn for each stem independently of other stems. The training mixture chunks were then recomputed from the random stem chunks. Testing was done on the entire track, using overlap-add in the same way as in [11], with a chunk size of 8.0 s and a hop size of 1.0 s.

Each model was trained using 8 NVIDIA A100 Tensor Core GPUs (40 GB each) on an Amazon EC2 p4d.24xlarge instance. Testing and inference were done using NVIDIA T4 Tensor Core GPUs on an Amazon EC2 g4dn.metal instance. An updated implementation of Bandit in this work employs significant gradient checkpointing to reduce the memory footprint, thus significantly increasing the chunk size and batch size per GPU memory, compared to [11].[6]

V. Results and Discussion

For brevity, we report only the median track-level SNRs in the paper. Mean SNR, mean SI-SNR, and median SI-SNR follow a similar trend to median SNR. Raw clip-wise SNR and SI-SNR are available in the repository. The median SNR results are provided in Tab. IV.

A. Monolingually Trained Models

To first demonstrate the issues of using monolingually trained models, we trained six models using monolingual variants of DnR v3 in ENG, DEU, FAO, FRA, SPA, and CMN. These languages were chosen as training variants as they consist of more than 30 hours of total data. We then test each model exhaustively on the test set of the aforementioned six monolingual variants, additional three monolingual test sets (EUS, JPN, and YOR), and three multilingual test sets grouped by language family (INC, DRA, and BNT).

The in-language DX performance exceeds 15 dB SNR in all six languages. Similarly, in-language MX and FX SNRs all exceed 10 dB and 9.5 dB, respectively. For the DX and FX stems, this is a similar result to the model trained on English-only DnR v2 as reported in [11]. The increase in MX SNR is likely due to the removal of vocals from DnR v2. For cross-language evaluation, however, the FAO- and CMN-trained variants struggled to generalize to other languages, performing at under 8 dB SNR for all but one test set each. The ENG-, DEU-, FRA-trained models appear to generalize significantly better, struggling slightly with FAO and more so with JPN test sets. The SPA-trained model appears to perform well on EUS, INC, DRA, and BNT test sets, but struggled with the rest.

It is important to note, however, that due to the various sources each language drew its data from, the performance variations cannot be solely explained by linguistic differences. Notably, ENG, DEU, FRA variants drew their data from crowd-sourced audiobook websites. As a result, their data contains significantly more diverse sets of acoustic environments and recording setups. Raw source data for SPA, EUS, INC, DRA, BNT, and YOR were all collected as a part of a Google initiative, thus likely share similar recording hardware and portable studio setups, albeit in different geographical locations [14, Sect. 2.2]. AISHELL-3, the source dataset for CMN, appears to have been recorded in a very uniform manner, likely in a single acoustic environment [29, Sect. 2]. Similarly, Faroese BLARK [17], the source dataset for FAO, was always recorded using a TASCAM DR-40, although in a few different acoustic environments [51]. Having a very uniform acoustic environment and recording setup is usually beneficial for TTS, the original purpose of these datasets. Our somewhat unintended use case

[6] We attempted to use mixed precision through multiple approaches, including only in parts of the model. However, none of the attempted approaches were sufficiently stable during training, thus were not employed in this work.

TABLE IV
PERFORMANCE OF BANDIT MODELS ON DIFFERENT VARIANTS OF DNR V3

	Dialogue SNR (dB), by Train Variant							Music SNR (dB), by Train Variant							Effects SNR (dB), by Train Variant						
Test Var.	MULTI	ENG	DEU	FAO	FRA	SPA	CMN	MULTI	ENG	DEU	FAO	FRA	SPA	CMN	MULTI	ENG	DEU	FAO	FRA	SPA	CMN
MULTI	**15.8**	15.1	14.9	2.5	14.9	7.4	3.2	10.3	10.3	**10.4**	8.6	10.3	9.0	8.7	**10.0**	9.5	9.6	–0.3	9.5	4.7	1.1
ENG	15.3	**15.6**	14.9	2.5	14.4	8.9	1.2	**10.4**	**10.4**	**10.4**	8.3	**10.4**	8.7	8.0	**9.9**	**9.9**	9.8	–0.1	9.3	6.7	–0.9
DEU	15.3	**15.4**	15.2	0.9	15.3	8.8	0.5	9.8	**10.0**	**10.0**	7.1	**10.0**	7.6	6.3	9.5	9.4	**9.6**	–3.2	9.5	5.9	–3.0
FAO	14.0	12.7	12.8	**15.1**	13.0	6.5	2.9	10.4	10.4	**10.5**	10.3	10.4	9.9	9.8	9.3	8.3	8.6	**10.2**	8.7	4.4	1.7
FRA	15.3	14.0	14.3	0.2	**15.7**	3.0	0.6	9.7	9.6	9.7	6.2	**10.0**	5.7	5.6	9.5	8.7	9.0	–3.7	**9.8**	0.2	–2.7
SPA	**16.2**	15.9	15.5	4.6	15.3	**16.2**	3.4	10.8	10.8	**10.9**	9.9	10.8	10.8	9.9	**10.6**	10.3	10.4	2.9	10.3	10.5	2.6
CMN	**15.5**	13.4	13.6	0.4	14.0	2.2	15.4	**10.6**	10.4	10.5	9.0	10.5	9.3	**10.6**	**10.3**	9.2	9.6	–1.4	9.6	0.5	10.2
EUS	**15.0**	14.2	13.9	2.3	14.0	13.8	6.3	10.3	10.3	**10.4**	9.3	**10.4**	9.8	9.7	**10.2**	9.7	9.8	0.3	9.7	9.5	4.7
JPN	**14.9**	9.0	8.9	2.5	9.7	3.9	0.6	**10.2**	10.0	10.1	8.6	10.1	9.6	9.2	**9.5**	5.5	5.9	0.1	6.3	1.2	–2.2
INC	16.8	**16.9**	16.5	2.6	16.2	15.6	2.1	10.7	10.8	**10.9**	9.3	10.8	10.5	9.6	10.5	10.4	**10.6**	0.2	10.4	10.0	0.0
DRA	16.9	**17.0**	16.6	6.2	16.2	16.4	3.0	10.8	10.8	**10.9**	9.8	10.8	10.6	9.7	10.6	10.5	**10.7**	3.7	10.4	10.3	1.1
BNT	**16.3**	14.3	14.1	12.1	14.3	12.3	5.8	**10.6**	10.3	10.5	10.0	10.4	10.0	9.9	**10.3**	9.0	9.1	8.1	9.1	8.1	3.8
YOR	**16.2**	14.4	14.6	7.1	15.6	9.3	9.7	**10.6**	10.3	10.5	9.8	**10.6**	9.8	10.1	**10.6**	9.9	10.1	5.1	10.5	6.7	7.6

Bold numbers indicate the best-performing model(s) by stem and testing variant.

in source separation might indicate that a diversity of acoustic environment and recording setup is an additional important factor for CASS model generalizability.

B. Multilingually Trained Model

Next, we investigate the performance of the multilingually trained model, which has training materials from all languages listed in Tab. II. Note that the number of hours of DX content seen by the model is identical (100 h) to monolingual models; the 100 h is now shared between 32 languages instead of being dedicated to one language.

The multilingual model generally performed well across all languages, with upwards of 14 dB SNR on DX, 9.7 dB SNR on MX, and 9.3 dB on FX. For the languages with dedicated training sets, the multilingual model performs either better or within 1.1 dB of their respective monolingual model on DX, 0.3 dB on MX, and 0.9 dB on FX. For the test sets without corresponding training sets, the multilingual model either performs the best or within 0.2 dB of the best-performing monolingual variants (either ENG or DEU) across all stems. Moreover, the multilingual model significantly outperforms all monolingual variants on JPN in the DX stem by at least 5 dB and in the FX stem by at least 3.5 dB, despite very limited JPN training materials. In general, this demonstrates that the multilingual model can consistently perform well across multiple languages, and is usually on par or better with dedicated monolingual models, even with comparatively less language-specific data to the language. This means that it is possible to only use one model to perform CASS in many languages without needing to train many dedicated models. This would particularly benefit low-resource languages with insufficient data to create a training set. Admittedly, the ENG and DEU also perform similarly or slightly better than the multilingual model on some test sets. However, these two monolingual models are less consistent than multilingual models in terms of their performance across languages.

VI. CONCLUSION

In this work, we present version 3 of the Divide and Remaster dataset, addressing some of the areas of improvement that have been identified in DnR v2. Specifically, DnR v3 tackled issues with regard to vocals and vocalization in music and effects stem, loudness distribution, mastering, and audio formats. Moreover, while DnR v2 is an English-only dataset, the dialogue stem in DnR v3 drew its content from 32 languages. Benchmarking experiments using Bandit demonstrated that a multilingually trained model can perform on par or close to dedicated monolingual models, enabling CASS on contents in languages with low data availability.

Despite our efforts, however, CASS remains in an early stage of research, with significant areas for future work both in the dataset curation and in the model development that are out of scope for this paper. Specifically, 1) a significant number of languages or even entire language families remain missing from DnR v3; 2) emotional diversity within the dialogue stem remains fairly limited in most languages; 3) spatialization, equalization, reverberation, and several other production aspects of cinematic contents remain unaddressed; and 4) nuances around the issues of non-speech human vocalized sounds continue to evade the strict three-stem setup. The authors are actively working to address these issues and welcome any effort to contribute to future iterations of the dataset.

ACKNOWLEDGMENT

The authors thank the following people and teams who have assisted with the project in various ways: Darius Pétermann; Gordon Wichern and the Mitsubishi Electric Research Laboratories (MERL) Speech & Audio team; Pablo Delgado and the Netflix Machine Learning Platform team; Scott Kramer and the Netflix Sound Technology team; Phillip A. Williams, William Wolcott, and the Netflix Audio Algorithms team.

Part of this work was done while K. N. Watcharasupat was separately supported by the IEEE Signal Processing Society Scholarship and the American Association of University Women (AAUW) International Fellowship.

ETHICS STATEMENT

CASS has emerging applications in improving content accessibility and internationalization. Within a production context, CASS was developed to *assist* human creatives with more tedious and repetitive aspects of cinematic audio production, in the hope that this would enable them to focus more of their time on the creative aspects of their work.

Divide and Remaster version 3 was developed using data derived exclusively from sources with explicit and permissive licenses, permitting commercial and derivative use. The dataset was developed in consultation with the lead authors of the original dataset and released under the same name with a major version change for continuity. The Bandit model employed is non-generative in nature; it can only filter audio content from its input and is incapable of outputting any audio content that was not already present within its input.

References

[1] I. Kavalerov, S. Wisdom, H. Erdogan *et al.*, "Universal Sound Separation," in *Proc. WASPAA*, Aug. 2019.

[2] C. Uhle, O. Hellmuth, and J. Weigel, "Speech enhancement of movie sound," in *Proc. AES Conv.*, 2008.

[3] J. Paulus, M. Torcoli, C. Uhle *et al.*, "Source Separation for Enabling Dialogue Enhancement in Object-based Broadcast with MPEG-H," *J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 510–521, Aug. 2019.

[4] D. Rieger, C. Simon, M. Torcoli *et al.*, "Dialogue Enhancement with MPEG-H Audio: An Update on Technology and Adoption," in *Proc. AES Conv.*, Helsinki, Finland, 2023.

[5] D. Petermann, G. Wichern, Z.-Q. Wang *et al.*, "The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks," in *Proc. ICASSP*, Singapore, Singapore, 2022.

[6] D. Petermann, G. Wichern, A. S. Subramanian *et al.*, "Tackling the Cocktail Fork Problem for Separation and Transcription of Real-World Soundtracks," *IEEE/ACM TASLP*, vol. 31, pp. 2592–2605, Dec. 2023.

[7] V. Panayotov, G. Chen, D. Povey *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210.

[8] M. Defferrard, K. Benzi, P. Vandergheynst *et al.*, "FMA: A Dataset for Music Analysis," in *Proc. ISMIR*, Suzhou, China, 2017, pp. 316–323.

[9] E. Fonseca, X. Favory, J. Pons *et al.*, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM TASLP*, vol. 30, pp. 829–852, 2022.

[10] S. Uhlich, G. Fabbro, M. Hirano *et al.*, "The Sound Demixing Challenge 2023 – Cinematic Demixing Track," *Trans. ISMIR*, Aug. 2023.

[11] K. N. Watcharasupat, C.-W. Wu, Y. Ding *et al.*, "A Generalized Bandsplit Neural Network for Cinematic Audio Source separation," *IEEE OJSP*, vol. 5, pp. 73–81, 2023.

[12] Y.-N. Hung, C.-W. Wu, I. Orife *et al.*, "A large TV dataset for speech and music activity detection," *EURASIP JASM*, vol. 2022, no. 1, p. 21, Sep. 2022.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.

[14] A. Butryna, S.-H. C. Chu, I. Demirsahin *et al.*, "Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview," in *Proc. LT4All*, Oct. 2020.

[15] I. Demirsahin, O. Kjartansson, A. Gutkin *et al.*, "Open-source Multi-speaker Corpora of the English Accents in the British Isles," in *Proc. LREC*, Marseille, France, May 2020, pp. 6532–6541.

[16] P. Puchtler, J. Wirth, and R. Peinl, "HUI-Audio-Corpus-German: A high quality TTS dataset," in *Proc. German Conf. AI*, Jun. 2021.

[17] A. Simonsen, S. S. Lamhauge, I. N. Debess *et al.*, "Creating a Basic Language Resource Kit for Faroese," in *Proc. LREC*, Marseille, France, Jun. 2022, pp. 4637–4643.

[18] S. Felice, S. V. Evain, S. Rossato *et al.*, "Audiocite.net : A Large Spoken Read Dataset in French," in *Proc. LREC-COLING*, Torino, Italia, May 2024, pp. 1795–1800.

[19] M.-G. Di Benedetto, S. Shattuck-Hufnagel, J.-Y. Choi *et al.*, "The LaMIT database: A read speech corpus for acoustic studies of the Italian language toward lexical access based on the detection of landmarks and other acoustic cues to features," *Data Brief*, vol. 42, p. 108275, Jun. 2022.

[20] O. Kjartansson, A. Gutkin, A. Butryna *et al.*, "Open-Source High Quality Speech Datasets for Basque, Catalan and Galician," in *Proc. SLTU/CCURL*, Marseille, France, May 2020, pp. 21–27.

[21] A. Guevara-Rukoz, I. Demirsahin, F. He *et al.*, "Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech," in *Proc. LREC*, Marseille, France, May 2020, pp. 6504–6513.

[22] N. Kadyrbek, M. Mansurova, A. Shomanov *et al.*, "The Development of a Kazakh Speech Recognition Model Using a Convolutional Neural Network with Fixed Character Level Filters," *Big Data Cogn. Comput.*, vol. 7, no. 3, p. 132, Sep. 2023.

[23] Y. Smoliakov, "Open Source Ukrainian Text-to-Speech datasets," 2023.

[24] N. Halabi, "Modern Standard Arabic Phonetics for Speech Synthesis," Ph.D. dissertation, University of Southampton, Jul. 2016.

[25] K. Sodimana, P. De Silva, S. Sarin *et al.*, "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in *Proc. SLTU*, Aug. 2018, pp. 66–70.

[26] F. He, S.-H. C. Chu, O. Kjartansson *et al.*, "Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems," in *Proc. LREC*, Marseille, France, May 2020, pp. 6494–6503.

[27] Z. Husein, "Malaya-Speech," Mesolitica, May 2024.

[28] Y. M. Oo, T. Wattanavekin, C. Li *et al.*, "Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech," in *Proc. LREC*, Marseille, France, May 2020, pp. 6328–6339.

[29] Y. Shi, H. Bu, X. Xu *et al.*, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech*, 2021, pp. 2756–2760.

[30] D. Xin, J. Jiang, S. Takamichi *et al.*, "JVNV: A Corpus of Japanese Emotional Speech with Verbal Content and Nonverbal Expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.

[31] J. Koguchi and S. Takamichi, "PJS: Phoneme-balanced Japanese singing voice corpus," Jun. 2020.

[32] D. V. Niekerk, C. V. Heerden, M. Davel *et al.*, "Rapid Development of TTS Corpora for Four South African Languages," in *Proc. Interspeech*, Aug. 2017, pp. 2178–2182.

[33] A. Gutkin, I. Demirşahin, O. Kjartansson *et al.*, "Developing an Open-Source Corpus of Yoruba Speech," in *Proc. Interspeech*, 2020, pp. 404–408.

[34] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 27th ed. SIL International, 2024.

[35] U. Gut and J.-T. Milde, "The prosody of Nigerian English," in *Proc. Int. Conf. Speech Prosody*, Apr. 2002, pp. 367–370.

[36] M. P. Barnes and E. Weyhe, "Faroese," in *The Germanic Languages*, 1st ed. Routledge, 1995, pp. 190–218.

[37] X. L. Regueira, "Galician," *J. Int. Phonetic Assoc.*, vol. 26, no. 2, pp. 119–122, Dec. 1996.

[38] L. A. Hitchcock, "A brief introduction to the sounds of Levantine Arabic," https://livingarabic.com/brief-introduction-to-the-sounds-of-levantine-arabic, 2020.

[39] P. Dasgupta, "Bangla," in *The Indo-Aryan Languages*. Routledge, 2007/0726, pp. 351–390.

[40] C. P. Masica, *The Indo-Aryan Languages*, ser. Cambridge Language Surveys. Cambridge: Cambridge University Press, 1993.

[41] R. Pandharipande, "Marathi," in *The Indo-Aryan Languages*. Routledge, Jul. 2007, pp. 698–728.

[42] V. Sok, *Basic Khmer*. LibreTexts, Nov. 2023.

[43] H. Abdullah, "The Morphology of Malay," Ph.D. dissertation, The University of Edinburgh, 1972.

[44] M. Jessen and J. C. Roux, "Voice quality differences associated with stops and clicks in Xhosa," *J. Phonetics*, vol. 30, no. 1, pp. 1–52, Jan. 2002.

[45] R. A. Paroz, *Elements of Southern Sotho*, 1st ed. Basutoland, Morija Sesuto Book Depot, 1946.

[46] W. G. Bennett, M. Diemer, J. Kerford *et al.*, "Setswana (South African)," *J. Int. Phonetic Assoc.*, vol. 46, no. 2, pp. 235–246, Aug. 2016.

[47] International Telecommunication Union, "ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level," Oct. 2015.

[48] C. J. Steinmetz and J. D. Reiss, "Pyloudnorm: A simple yet flexible loudness meter in Python," in *Proc. AES Conv.*, 2021.

[49] European Broadcasting Union, "EBU R 128-2023: Loudness normalisation and maximum level of audio signals," 2023.

[50] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, May 2019.

[51] I. N. Debess, "Sound recordings of project Ravnur," Feb. 2022.