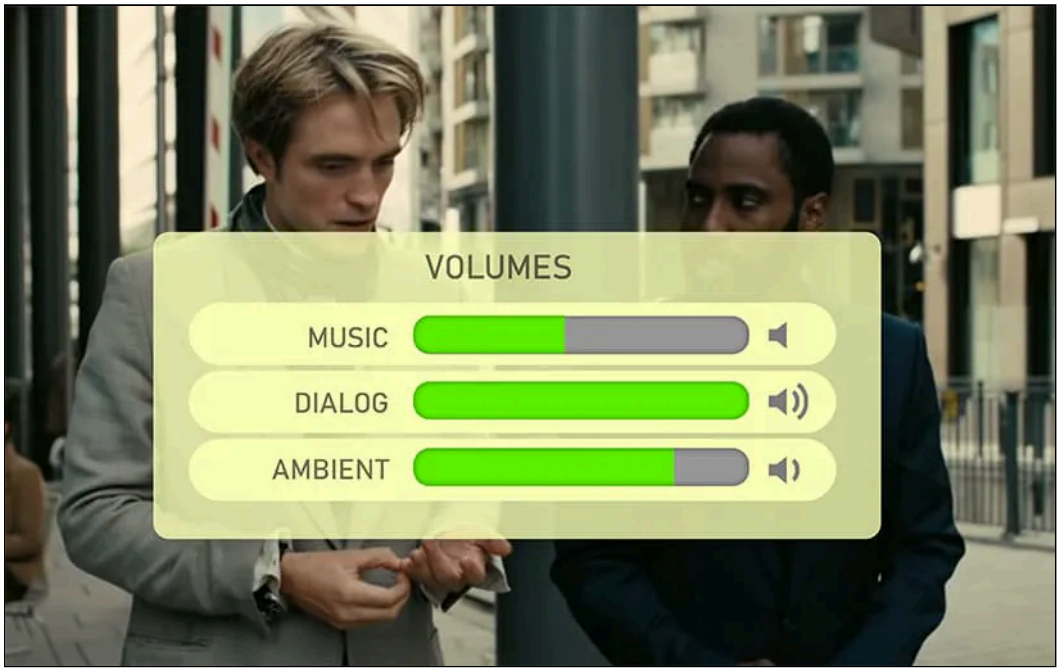
AI Research Envisages Separate Volume Controls for Dialog, Music and Sound Effects

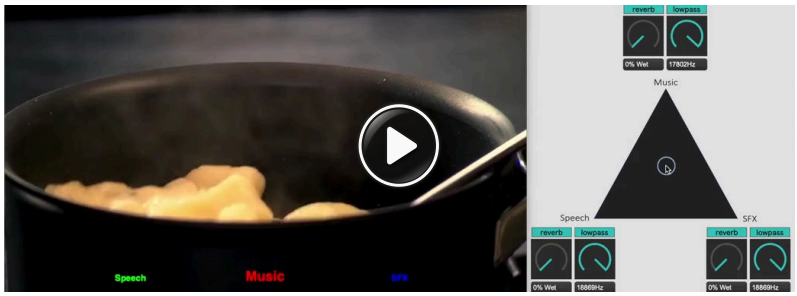
Originally published at https://www.unite.ai/ai-research-envisages-separate-volume-controls-for-dialog-music-and-sound-effects/



A new research collaboration led by Mitsubishi investigates the possibility of extracting three separate soundtracks from an original audio source, breaking down the audio track into speech, music and sound effects (i.e. ambient noise).

Since this is a post-facto processing framework, it offers potential for later generations of multimedia viewing platforms, including consumer equipment, to offer three-point volume controls, allowing the user to raise the volume of dialog, or lower the volume of a soundtrack.

In the short clip below from the accompanying video for the research (see end of article for full video), we see different facets of the soundtrack being emphasized as the user drags a control across a triangle with each of the three audio components in one corner:



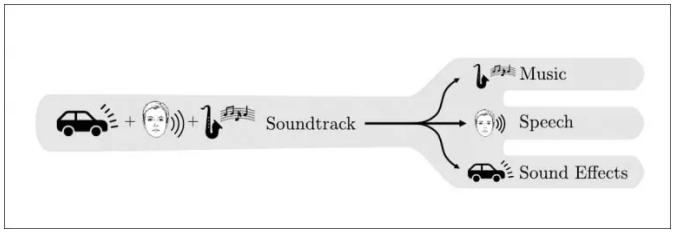
CLICK TO PLAY VIDEO ON SITE

A short clip from the video accompanying the paper (see embed at end of article). As the user drags the cursor towards one of the three extracted facets in the triangle UI (on the right), the audio emphasizes that part of the tripartite soundtrack. Though the longer video cites a number of additional examples on YouTube, these seem currently to be unavailable. Source: https://vimeo.com/634073402

The paper is entitled *The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks*, and comes from researchers at the Mitsubishi Electric Research Laboratories (MERL) in Cambridge, MA, and the Department of Intelligent Systems Engineering at Indiana University in Illinois.

Separating Facets of a Soundtrack

The researchers have dubbed the challenge 'The Cocktail Party Problem' because it involves isolating severely enmeshed elements of a soundtrack, which creates a roadmap resembling a fork (see image below). In practice, multi-channel (i.e. stereo and more) soundtracks may have differing amounts of types of content, such as dialog, music, and ambience, particularly since dialog tends to dominate the center channel in Dolby 5.1 mixes. At present, however. the very active research field of audio separation is concentrating on capturing these strands from a single, baked soundtrack, as does the current research.



The Cocktail Fork – deriving three distinct soundtracks from a merged and single soundtrack. Source: https://arxiv.org/pdf/2110.09958.pdf

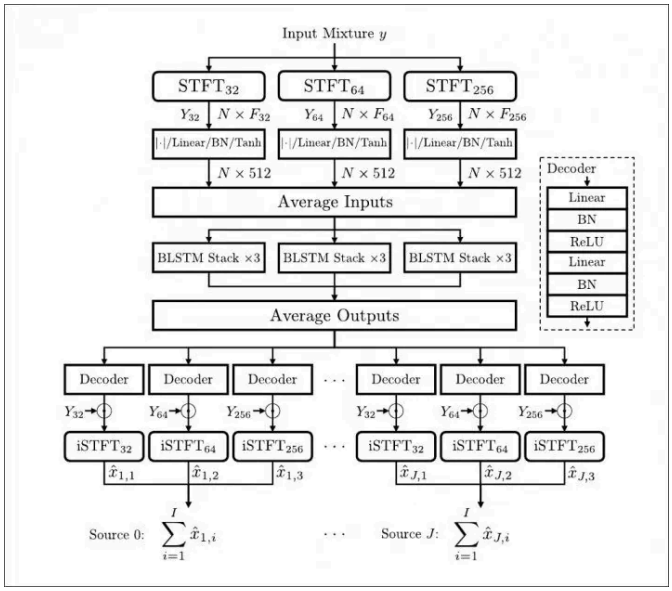
Recent research has concentrated on extracting speech in various environments, often for purposes of denoising speech audio for subsequent engagement with Natural Language Processing (NLP) systems, but also on the isolation of archival singing voices, either to create synthetic versions of real (even dead) singers, or to facilitate Karaoke-style music isolation.

A Dataset for Each Facet

To date, little consideration has been given to using this kind of AI technology to give users more control over the mix of a soundtrack. Therefore the researchers have formalized the problem and generated a new dataset as an aide to ongoing research into multi-type soundtrack separation, as well as testing it on various existing audio separation frameworks.

The new dataset that the authors have developed is called Divide and Remaster (DnR), and is derived from prior datasets LibriSpeech, Free Music Archive and the Freesound Dataset 50k (FSD50K). For those wishing to work with DnR from scratch, the dataset must be reconstructed from the three sources; otherwise it will shortly be made available at Zenodo, the authors claim. However, at the time of writing, the provided GitHub link for source extraction utilities is not currently active, so those interested may need to wait a while.

The researchers have found that the CrossNet un-mix (XUMX) architecture proposed by Sony in May in works particularly well with DnR.



Sony's CrossNet architecture.

The authors claim that their machine learning extraction models work well on soundtracks from YouTube, though the evaluations presented in the paper are based on synthetic data, and the supplied main supporting video (embedded below) is currently the only one that seems to be available.

The three datasets used each comprise a collection of the kind of output that needs to be separated out from a soundtrack: FSD50K is occupied with sound effects, and features 50,000 44.1 kHz mono audio clips tagged with 200 class labels from Google's AudioSet ontology; the Free Music Archive features 100,000 stereo songs covering 161 music genres, though the authors have used a subset containing 25,000 songs, for parity with FSD50K; and LibriSpeech provides DnR with 100 hours of audio book samples as 44.1kHz mp3 audio files.

Future Work

The authors anticipate further work on the dataset and a combination of the separate models developed for additional research into speech recognition and sound classification frameworks, featuring automatic caption generation for speech and non-speech sounds. They also intend to evaluate possibilities for remixing approaches that can reduce perceptual artifacts, which remains the central problem when dividing a merged audio soundtrack into its constituent components.

This kind of separation could in the future be available as a consumer commodity in smart TVs that incorporate highly optimized inference networks, though it seems likely that early implementations would need some level of pre-processing time and storage space. Samsung already uses local neural networks for upscaling, while Sony's Cognitive Processor XR, used in the company's Bravia range, analyzes and reinterprets soundtracks on a live basis via lightweight integrated AI.

Calls for greater control over the mix of a soundtrack recur periodically, and most of the solutions offered have to deal with the fact that the soundtrack has already been bounced down in accordance with current standards (and presumptions about what viewers want) in the movie and TV industries.

One viewer, vexed at the shocking disparity of volume levels between various elements of movie soundtracks, became desperate enough to develop a hardware-based automatic volume adjuster capable of equalizing volume for movies and TV.

Though smart TVs offer a diverse range of methods to attempt to boost dialog volume against grandiose volume levels for music, they're all struggling against the decisions made at mixing time, and, arguably, the visions of content producers that wish the audience to experience their soundtracks exactly as they were set up.

Content producers seem likely to rankle against this potential addition to ‘remix culture’, since several industry luminaries have already voiced discontent against default post-processing TV-based algorithms such as motion smoothing.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More