Audio Engineering Society

Convention Paper 10427

Presented at the 149th Convention
Online, 2020 October 27-30

Dialog Enhancement via Spatio-Level Filtering and Classification

Aaron S. Master[1], Lie Lu[1], Heidi-Maria Lehtonen[2], Harald Mundt[3], Heiko Purnhagen[2] and Daniel P. Darcy[1]

[1] *Dolby Laboratories, Inc, 1275 Market St, San Francisco, CA 94103, United States*

[2] *Dolby Sweden AB, Gävlegatan 12A, 113 30 Stockholm, Sweden*

[3]*Dolby Germany GmbH, Deutschherrnstraße 15, 90429 Nürnberg, Germany*

Correspondence should be addressed to Aaron S Master (Aaron.Master@dolby.com)

ABSTRACT

Dialog enhancement (DE) is a feature that allows a listener to increase the level of dialog in a content item relative to backgrounds. DE is "unguided" if only the finished mix is available, meaning that a DE system must estimate the dialog. Spatio-Level Filtering (SLF) is a source separation technology that, when combined with dialog classification, allows for high-quality unguided DE for typical entertainment content in a stereo or higher channel count format. SLF exploits spatial and level information and requires little lookahead, memory, computation and training data. To evaluate results, we conduct two subjective listening experiments which indicate favorable performance.

1 Introduction

When experiencing typical television or movie content, listeners may struggle to understanding spoken dialog [1], simply prefer to change the relative dialog level [2], or both. Dialog enhancement (DE) is a feature which allows users to increase the level of dialog relative to other sounds, typically referred to as backgrounds or music and effects ("M&E"). Typical backgrounds include music, crowd noise, "walla" (crowd din) and special effect sounds including Foley. If a DE system is supplied with the original dialog and background (or complete mix) tracks, it performs *guided DE*. A trivial process allows relative dialog level to be increased: the system increases the level of the supplied dialog track, decreases the supplied backgrounds, or both. The case where dialog level is increased may be expressed as:

$$y = gd + x \qquad (1)$$

where x is the original mix signal (consisting of dialog d plus backgrounds b), g is the dialog boost factor derived from listener input, and y is the *ideal dialog-boosted signal*. That is, $y = (g + 1)d + b$. This DE method is called "boost type" because the dialog level is increased in y relative to the input signal x while the backgrounds remain at their original level. The boost gain g may be derived from the desired decibel increase in dialog g_{dB} via $g = 10\text{^}(g_{\mathrm{dB}}/20) - 1$. Scaling may also be applied to y such that the boosted signal preserves the loudness or level of d or y. For simplicity, this work will assume boost type DE only.

If only the complete mix is available (no original dialog track), then the problem is more challenging; in order to boost dialog, a system must first estimate it via source separation. (For a general review, see e.g. [3, 4].) This problem is now termed *unguided DE*. Equation (1) becomes

$$y = g\hat{d} + x \quad (2)$$

where $\hat{d}$ is estimated dialog and y is now termed the *estimated boosted signal*. This is shown in Figure 1. Such problems are generally made easier with more input channels, as in 5.1 or higher channel count audio (see, e.g. [3]). For typical delivery, however, stereo (2 channel) input is common, and in some cases will be the most common format to which a DE system has access. The number of background sounds is generally unknown and could be larger than one, leading to an underdetermined source separation problem (2 channels, more than 2 sources including dialog). It can be made yet more challenging in practical application to entertainment content, because computation, memory and lookahead may be limited, depending on where the DE system exists in the creation and delivery chain. In this work, we choose to focus on the stereo case because it is common, challenging, and can lead to solutions which can be adapted to higher channel-count cases.

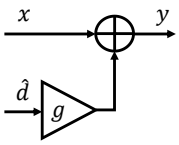

Figure 1: Estimated boosted signal flow

To address these challenges and estimate $\hat{d}$, we introduce a system which combines spatio-level filtering (SLF) source separation with dialog classification. SLF uses the target source signal model and features of [5] which was developed for stereo music, but with a more general model of backgrounds which simplifies the training process. SLF extracts signals whose spatial and level characteristics make them good dialog candidates, and a classifier estimates whether the extracted signals (or the original mix) contain dialog. The classification information is used to gate the SLF output, thereby producing an estimated dialog signal $\hat{d}$. This is shown in Figure 2

DE as defined here substantively differs from other audio signal processing goals, such as eliminating backgrounds or modifying speech sounds (possibly by changing timbre) to increase intelligibility. Such

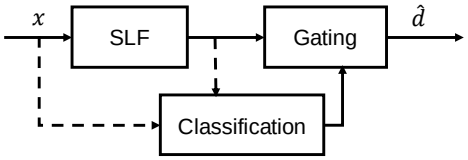

Figure 2: SLF + classification signal flow.

efforts may be termed noise reduction (see e.g. [6]) or speech enhancement (see e.g. [7]), but they are sometimes also called "dialog enhancement." For the present unguided DE task, the output signals deliberately include backgrounds, and the system aims to prevent any distortion of backgrounds while modifying dialog only by changing its relative level. This impacts both signal processing goals and system evaluation.

The SLF source separation system can be optimized for various audio applications. The version described here is designed to work with a classifier to extract dialog in entertainment content, whether it is center-panned, non-center-panned, mixed with delay, phase-modified or reverberant. This makes it more general than a system which extracts only center-panned dialog (e.g. [8]). To do this efficiently, the system uses an adaptive approach termed "Shift and Squeeze" (S&S). SLF requires little lookahead, memory, computation and training data. These characteristics make the "SLF + Classification" (SLF+C) system suitable for various applications, including DE at encode, within cloud-based media workflows, or on an endpoint device.

This paper is organized as follows. In section 2, we describe the SLF model and related concepts. In section 3, we motivate detection of spatially identifiable sources and use of the adaptive system. In section 4, we describe SLF system operation in detail, including perceptual optimizations. In section 5, we describe how classification data is combined with the SLF system output to produce a dialog estimate within latency constraints. In section 6, we introduce the concept of dialog boost quality, and provide related results from human listening tests. We conclude in section 7 with a summary and discussion of future work.

2 SLF Model

While the SLF system described herein is designed to extract any spatially identifiable sources, it is based on a core Bayesian SLF system which models and extracts panned sources. This section defines the corresponding signal model, context, features and concepts. In so doing, we develop an intuition for the utility of the information exploited by SLF.

2.1 Signal Model

This model assumes basic time domain mixing of a target source s_1 and backgrounds into two channels, termed "left" (x_1 or X_1) and "right" (x_2 or X_2). The target source shall be assumed to be amplitude panned using the constant power law:

$$\begin{aligned} x_1 &= \cos(\Theta_1)\, s_1 \\ x_2 &= \sin(\Theta_1)\, s_1 \end{aligned} \tag{3}$$

where Θ_1 ranges from 0 (far left) to $\pi/2$ (far right). Because the STFT is linear, we may express this in the STFT domain as:

$$\begin{aligned} X_1 &= \cos(\Theta_1)\, S_1 \\ X_2 &= \sin(\Theta_1)\, S_1 \end{aligned} \tag{4}$$

where values exist for each bin ω and frame t, e.g. $X_1(\omega, t)$. Continuing in the STFT domain, we express addition of backgrounds B to each channel:

$$\begin{aligned} X_1 &= \cos(\Theta_1)\, S_1 + \cos(\Theta_B)\, |B| e^{j\angle B} \\ X_2 &= \sin(\Theta_1)\, S_1 + \sin(\Theta_B)\, |B| e^{j\angle B + \phi B} \end{aligned} \tag{5}$$

Here, in addition to a panning parameter Θ_B, the backgrounds B have other parameters $\angle B$ and ϕ_B. These parameters respectively describe the phase difference between S_1 and the left channel phase of B, and the interchannel phase difference for B only. (There is no need to include a ϕ_{S1} parameter because the interchannel phase difference for a panned source is by definition zero.) Target and backgrounds are assumed to share no phase relationship in STFT space, so we will model the distribution on $\angle B$ as uniform.

One can think of Θ_1 being a specific single value (the "panning parameter" for the target source) which completely specifies its mixing. There is a distribution on its level $|S|$, which we shall assume is approximately known, at least over roughly-octave subbands.

The background spatial parameters Θ_B and ϕ_B, respectively panning and interchannel phase difference, are understood to have a *distribution*. It is common for backgrounds to be diffuse, which manifests as widely distributed values for Θ_B and ϕ_B. As with the target source, there is also a distribution on the background level $|B|$ which we shall assume is known at least over roughly-octave subbands. For purposes of creating training data, these parameters represent all backgrounds, whether they are diffuse or concentrated such as in the case of panned background sources. This considerably simplifies the Bayesian training process over [5] in which backgrounds are assumed to be some number of panned sources, each combination of which has a Bayesian prior.

For purposes of this model, the source and backgrounds shall only be considered at points in time where both are assumed to be "active," meaning that both are present in the mix signal. A classifier described below reduces spurious source extraction in the absence of a target source.

2.2 Features

The SLF system operates in the STFT domain and takes only X_1 and X_2 as input. From these it calculates SLF features for each ω and t:

$$\begin{aligned} \theta &= \arctan\left(\frac{X_2}{X_1}\right) \\ \phi &= \angle\left(\frac{X_1}{X_2}\right) \\ U_{\mathrm{dB}} &= 10\log_{10}(|X_1|^2 + |X_2|^2) \end{aligned} \tag{6}$$

The first parameter, θ, is *detected panning* for each (ω, t) tile. It can be seen that if a panned source is dominant in a given tile (i.e. much higher in level than backgrounds), the detected panning will equal its panning parameter Θ_i.

The second parameter, ϕ, is detected interchannel phase difference for each tile on the range from $-\pi$ to π radians. (To prevent concentration detection

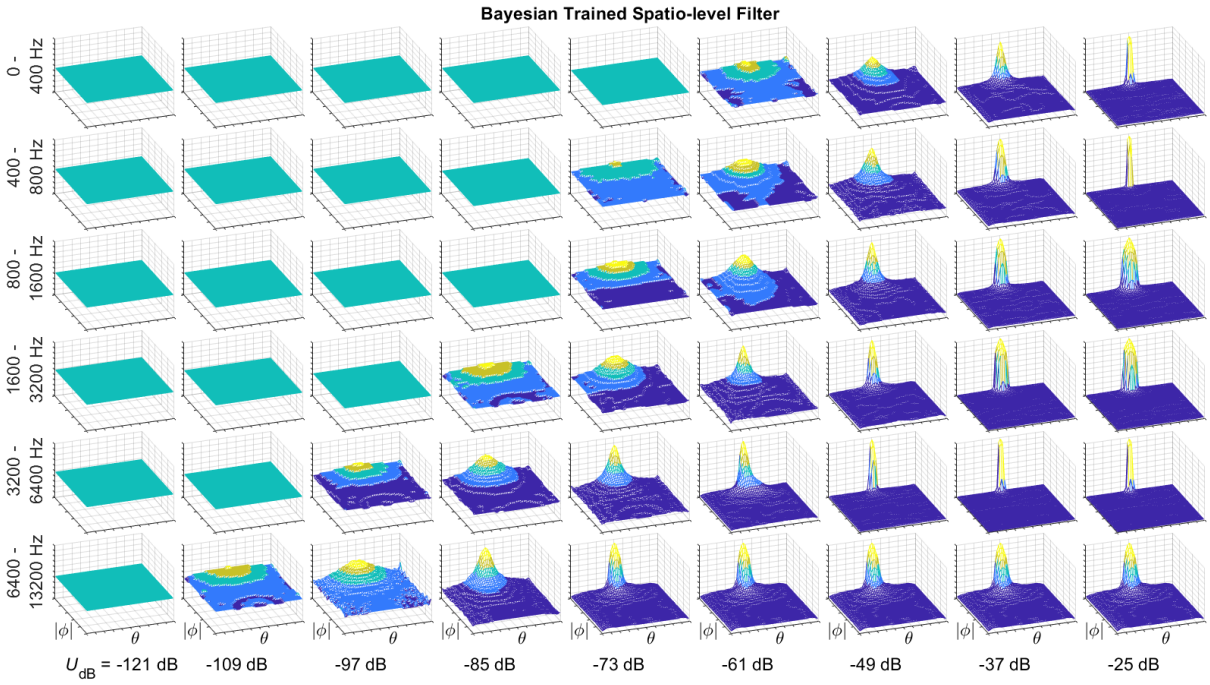

Fig. 3. A sampled representation of a Bayesian trained SLF. The four input variables are depicted as the left-right (θ) and in-out ($|\phi|$) axes of each subplot, and the subplot rows (b) and columns (U). The softmask output value is the vertical axis of each subplot. A sampling of U subplots is shown to allow reasonable figure width

bias, we also define $\phi 2$ as the identical data on the range from 0 to 2π.) If a panned source is dominant in a given tile, ϕ and $\phi 2$ will both be zero.

The third parameter, U, is detected level for each tile which is the dB version of the "Pythagorean" magnitude of the two channels. It may be thought of as a mono magnitude spectrogram. Various scalings of U may also be used, for example: $U_{\text{power}} = |X_1|^2 + |X_2|^2$.

2.3 Subbands and Chunks

To facilitate mixing parameter detection and other operations, the system groups frequency bins ω within quasi-octave subbands with band edges of 0, 400, 800, 1600, 3200, 6400, 13200, and 24000 Hz. For dialog extraction, the system will typically not process the highest subband. We denote the subband index b.

For mixing parameter detection, the system will also use *chunks* which are overlapping groups of consecutive frames. We use chunks of 10 frames (1 current, 4 lookahead, 5 lookback, a total buffer of about 277ms) with a chunk hop size (stride) of 5 frames. These chunk parameters can be modified per application; they balance parameter estimation stability, computation, responsiveness and latency.

2.4 Example Distributions and Filter

To develop an intuition for the SLF features, we briefly describe examples of U_{power} weighted 2-D distributions on θ and ϕ within subbands and chunks, which can be estimated via histograms.

- For a typical center-panned dialog source over quiet, diffuse backgrounds, there is a spike at $(\theta, \phi) = (\pi/4, 0)$, and quasi-flat lower values elsewhere.
- For a reverberant central dialog source, there is a less sharp peak also at $(\theta, \phi) = (\pi/4, 0)$.

- For a far-left panned source only, there are high values (uniform distribution) for all ϕ, at $\theta = 0$. For a far right source, the distribution is similar but at $\theta = \pi/2$. This is due to a lack of matching phase information in the opposite channel.
- For diffuse backgrounds only, there is a quasi-uniform (flat) distribution across (θ, ϕ).

The SLF system is trained to exploit such spatial differences found in its training data, as well as joint dependencies on level information.

To make this concept more concrete, we include a visualization of an example spatio-level filter trained to extract a center-panned source in the presence of moderately diffuse backgrounds, in Fig. 3. The filter outputs softmask values m between 0 and 1 which aim to exactly match the fraction of input energy belonging to the center-panned source: higher values mean the filter predicted more input energy belongs to the target. The filter takes a four-dimensional input for each ω and t, consisting of the corresponding (θ, ϕ, U, b) values. (Note that the subband parameter b is a trivial lookup of the subband to which ω belongs.) The figure shows the inputs and outputs on a single plot. Consider an example input tile which exists in frequency subband 4, at a U level of -61 dB, where θ is $\pi/4$ and ϕ is 0. (This corresponds to the middle of the back wall on the 6th horizontal and 4th vertical subplot.) It can be seen that about 80% of the input energy would be passed to the output. The trends seen in the plot are a function of the Bayesian training process. The training applied here follows the steps described in [5], but with the background model specified per the parameters in Eq. (5). For stability, and to balance artifacts and interferers, a 25th percentile of each distribution is used rather than an expected median value or most likely median. The filter may be trained and implemented via any variety of means, provided that the inputs and outputs are as specified.

3 Spatial Parameter Investigation

The system described below was built to exploit information found from investigations of typical examples of stereo television and movie content which contained dialog. We summarize three key findings. First, spatial concentration of energy within chunks and subbands correlates with intelligible dialog sources, even if those concentrations are not centered around $\phi = 0$. Next, detecting ϕ concentrations *within quasi-octave subbands* can be sufficient for identifying and extracting sources mixed with delay, even without explicitly estimating delay. This greatly simplifies delay-mixed source extraction, because it is considerably less challenging to detect interchannel phase difference (ϕ) concentrations than it is to reliably estimate interchannel delay. Third, for typical dialog extraction from entertainment content, it is effective and efficient to extract one source per frequency subband per unit time, and model subband sources as belonging to the same target source. Tracking more sources than this led to little perceptual benefit while substantially increasing complexity.

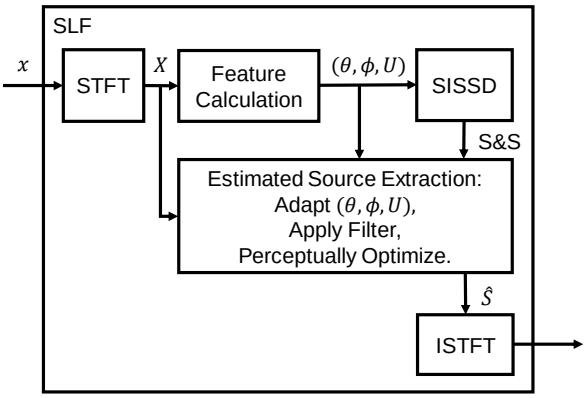

Figure 4: SLF System Operation

4 System Operation

The SLF system exists within the context of STFT domain softmask source separation systems (see e.g. [5]) which include four basic steps: (1) Apply STFT to each channel. (2) Detect the existence and mixing parameter(s) of target source(s). (3) Use the mixing parameters to extract estimated sources. (4) Invert the STFT domain representations to obtain stereo time domain estimate(s) of the target source(s). In subsections below, we describe how the SLF system performs each step. Step 2 includes a spatially identifiable subband source detector (SISSD) which estimates target sources' mixing parameters, even if

they do not correspond to a panned source. These parameters are used to adapt data such that, in step 3, an SLF built to extract center-panned sources can be used for arbitrary spatially identifiable sources. Step 3 differs from magnitude-only softmask systems, because it adds phase and panning optimization, which aim to perceptually improve results. Figure 4 is a flow diagram of the steps in this section.

4.1 STFT Front End and Feature Calculation

For the STFT front end, we assume, or convert inputs to have, a 48 kHz sample rate, and use 4096 sample frames with a square root of Hann window, hopped at 1024 samples (75% overlap). The corresponding 85.3 ms window length falls within the optimal range for monoresolution STFT speech separation systems as found by various methods (see e.g. [9, 10]). The 75% overlap represents a tradeoff between the minimum of 50% needed for perfect reconstruction and arbitrarily larger perfect reconstruction values which asymptotically improve quality [11] while exponentially increasing computation.

The system calculates (θ, ϕ, U) for each (ω, t) tile, and if necessary, adjusts U to match the level of the training data using a long term LKFS measurement [12].

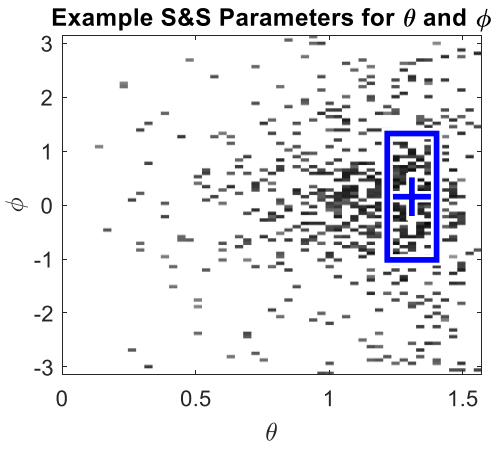

Fig. 5. Example 2-D heat map histogram with detected S&S parameters.

4.2 SISSD

The system next detects one spatially identifiable source per chunk per subband and characterizes it through "shift and squeeze" parameters.

For each chunk and subband within it, the system creates a U_{power} weighted 51-bin histogram on θ. It does the same for ϕ and $\phi 2$ but with 102 bins. The histograms are each smoothed over their given dimensions and vs chunks. For the smoothed θ histograms, the system detects the theta value of the highest peak, which we call thetaMiddle, and also the width around this peak necessary to capture 40% of energy in the histogram, which we call thetaWidth. It does the same for ϕ and $\phi 2$, recording phiMiddle, phi2Middle, phiWidth and phi2Width, but requiring 80% energy capture for widths. The system records final values for phiMiddle and phiWidth based on which had a higher concentration in phi space as indicated by a smaller phiWidth value.

An example 2-D θ and ϕ histogram heat map for one chunk and subband is shown in Fig. 5 (though parameters are calculated on 1-D histograms on θ and ϕ). Darker areas represent greater intensity. The width and height of the rectangular box overlaid on the histogram corresponds to the detected thetaWidth and phiWidth, and the "+" icon corresponds to thetaMiddle and phiMiddle.

The system converts thetaMiddle, thetaWidth and phiWidth to per-frame values via first order linear interpolation, and phiMiddle to per-frame values by zeroth order hold, to avoid rapid phase change when different ϕ ranges are chosen in different chunks. We term thetaMiddle and thetaWidth the "θ shift and squeeze" parameters, and phiMiddle and phiWidth the "ϕ shift and squeeze" parameters. Collectively they are "Shift and Squeeze" or "S&S" parameters.

Fig. 6 shows the S&S parameters versus chunk and subband for a sci-fi movie audio except where the dialog is bandlimited "radio voice"; initially mostly-left-panned dialog, followed by center-panned dialog. Observe that for dialog segments and subbands (2 through 4), phiMiddle, phiWidth, and thetaWidth are all near zero, as expected for panned sources, while

thetaMiddle is initially 0.1π, then $\pi/4$, as expected for mostly-left then center-panned sources. For bands lacking dialog, we observe greater values of phiWidth and thetaWidth and more random values of both thetaMiddle and phiMiddle, indicating more diffuseness. For sources with reverberation, thetaWidth and phiWidth are both typically larger than for panned sources. For sources mixed with delay, various phiMiddle values are typically seen in each subband, but theta values are more consistent.

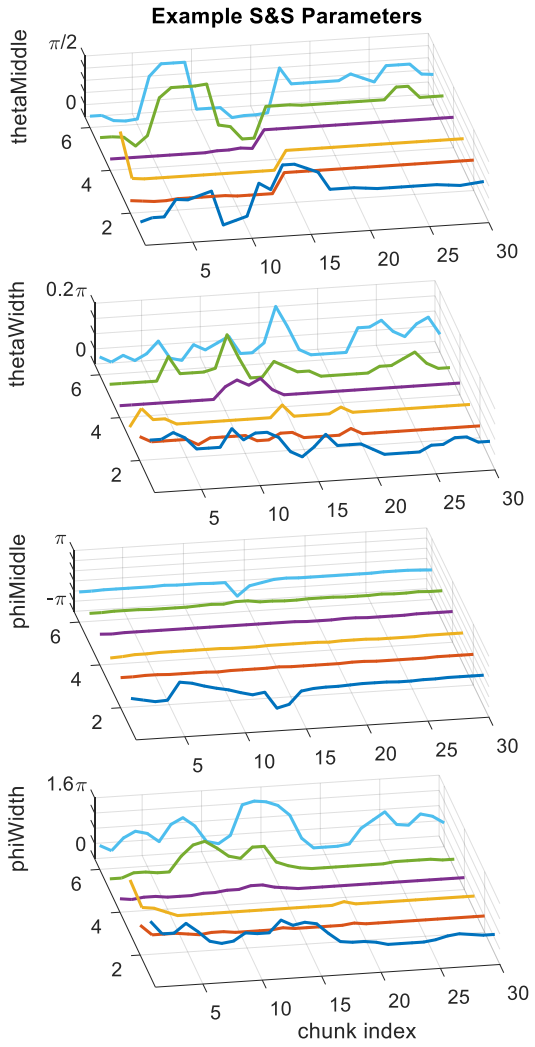

Fig. 6. Example S&S parameters vs chunk and subband.

4.3 Estimated Source Extraction

For the third stage, the system extracts an estimated target source in each subband and frame by using the S&S parameters to adapt the (θ, ϕ) parameters before applying a spatio-level filter which takes input of (θ, ϕ, U, b) for each bin and frame. Because the filter is trained on a center-panned source, the concept is to relocate the θ and ϕ data to be centered around $\theta = \pi/4$ (center) and $\phi = 0$ (panned). To do this, the system compares the S&S parameters to those for the center panned source data on which it was trained. If performs shift based on the "Middle" parameters and squeeze based on the "Width" parameters. To prevent spurious emphasis of spatially unconcentrated sources, there is a limit of 1.5x on the squeeze adaptations. (This is relaxed for single-channel extreme panned sources which lack ϕ concentration.) Adaptation occurs within each frame and subband, since this is the granularity of the S&S parameters.

This technique drastically reduces the amount of memory and computation needed for source extraction, because the system can use a single trained filter to extract a quasi-infinite number of spatially identifiable sources, rather than requiring a quasi-infinite number of trained filters. Fig. 7 is a stylized depiction of S&S parameters modifying example spatial characteristics of the input.

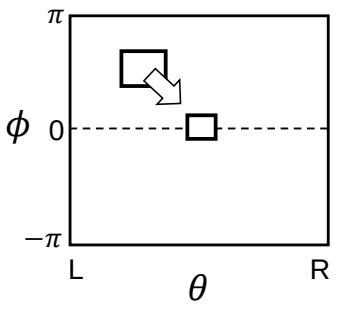

Fig. 7. Conceptual illustration of S&S.

4.3.1 Perceptual Optimizations

Instead of using the basic target source estimate described just above, the SLF system next optimizes phase, panning, and mask level for perceptual benefit. The STFT domain target source estimate obtained thus far uses trivial phase information copied from the input mix, a solution known to be suboptimal. It has been shown that even a rough estimate of phase can substantially improve source separation quality [13] or alter the perception of the volume level of a target source [5]. Presently, most estimated extracted sources are expected to be panned, which means that ϕ should be zero for all phase values; yet this will not universally apply to the source estimate when using input phase. The SLF system remedies this by requiring output ϕ to be zero. (Although this does not always match phiMiddle for delayed and reverberant sources, no negative effects were observed from this requirement, which effectively models estimated source phase as coming from a panned source.) To achieve this, the system performs *phase optimization* by applying a weighted average of the left and right channel phase to each channel.

A similar concept applies to the output θ values: if the system applies the same magnitude softmask to both channels, the resulting detected panning value will not always be thetaMiddle, the detected concentration. The system applies *panning optimization* by multiplying each left channel softmask value by cos(thetaMiddle) and each right channel value by sin(thetaMiddle), which results in source estimates whose θ values equal thetaMiddle for their frame and subband.

The effect of panning and phase optimization is that they allow sufficiently loud target source estimates to spatially mask interference, which has led listeners to describe resulting target source audio as "louder" or "clearer." Automated metrics will not necessarily reflect these benefits.

The magnitude mask levels themselves can also be perceptually optimized before smoothing. The Bayesian SLF filter used was trained to produce a solution which balances low interference and low artifacts in a target source estimate. However, for the DE case, SLF target source estimate will be added to the original mix which tends to mask artifacts. To shift this balance even after the filter has been trained, the system performs *bulk reduction*, in which softmask values below an interference-correlated *balance point* are reduced by a multiplicative scale factor. Here we chose a balance point of 0.51 and a scale factor of 0.33. These values should be optimized in accordance with the application and expected boost level range.

4.4 Invert STFT Representation

Finally, an inverse STFT with the same synthesis window as analysis window, is performed on each channel representation.

5 Classification, Gating and Latency

Above, we described how the SLF system outputs a candidate dialog signal which shall be gated based on a dialog classifier. We describe two options for classifier audio input as shown with dashed lines in the system diagram (Figure 2). First, the estimated SLF target source signal can be input to the classifier. This could lead to more accurate classification but requires the classifier to wait for SLF system output, which increases latency. A second option is to input the original mix directly to the classifier; latency is then the greater of SLF and classification latencies, rather than their sum. The present SLF system was designed for low-latency applications and uses the second option.

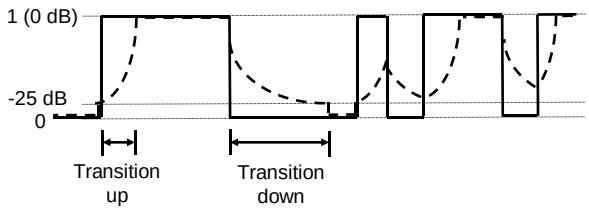

Fig. 8. Sample-level classification (solid) and gating function with causal transitions (dashed). Figure is for illustrative purposes; axes not to scale.

The SLF system has algorithmic latency of 470 ms resulting from use of (and smoothing across future) frames and chunks. If we allow for moderate or high

risk to audio quality, different sets of lookahead choices lead to 320 and 85 ms algorithmic latency, respectively. The classifier used has an algorithmic latency of approximately 700 ms. In applications where video and audio are delivered together, it may be costly or impossible for audio processing to require substantial delay from the video processing components. Additional latency is currently of little benefit to SLF because it exploits training information on a very short timescale.

The classifier itself uses features described in [14, 15] as inputs to a deep neural network, which is trained to output a confidence level between 0 and 1 per classifier frame, each 1024 samples in hop size. Choosing a trigger level of 0.45 leads to the lowest overall error rate (2.3% each for false negatives and positives). However, the target source candidates output by the SLF system are more robust to false positives than negatives (SLF tends to attenuate non-dialog) which leads us to choose a trigger level of 0.1, for a higher false positive rate of 7.0% and low false negative rate of 0.6%. Triggering frames are labeled 1 and non-triggering frames are labeled 0.

Next, we convert from frame level to sample level classification using zero order hold, then add causal transition regions which go up from -25 dB to 0 dB over 180 ms, and down over 800 ms. See Fig. 8 for an example. This time domain gating function is multiplied with the time domain SLF target source estimate for each channel to produce the estimated dialog signal $\hat{d}$.

6 Evaluation

6.1 Dialog Boost Quality

We now seek to evaluate the signal y as described in Eq. (2), in terms of practical system performance on representative examples at representative dialog boost levels. We will assess *dialog boost quality* or DBQ. To do so, we consider boosted signals including backgrounds, not dialog estimate signals, as they are not surfaced to the user. DBQ shall be described for a specific boost level in decibels, e.g. "9 dB DBQ." Achieving high DBQ is generally considered easier than achieving high quality source separation or complete noise suppression, because artifacts or interference that would be exposed in an estimated dialog signal can be masked when that signal is added back to the original mix for the boost case; this "mix masking" is greatest for low dialog boost levels and least for high levels. Results here will evaluate target boost levels of 9 and 15 dB, which are greater than or equal to levels commonly chosen by human listeners for typical entertainment content (see e.g. [2]).

Characteristic	Categories (Abbreviation)
Speaker gender presentation	Male only (M), female only (F), both male and female (B). Unless noted, dialog is from speakers perceived to be adult humans.
Background music	Large orchestral and or choral ensemble (L), small acoustic ensemble (A), synth-heavy pop (S), none (X).
Background effects	Crowd noise (C), race track sounds (R), mechanical sounds (M), spatial objects (S), ambient nature sounds (N), none (X).
Genre	Sports discussion (D), Product Ad (P), News documentary (N), Live motorsports (M), Live sports (L), Other TV / movie (T)

Table 1: Characteristics of audio items used in listening tests, with abbreviations.

In DE, the goal is generally to boost dialog without boosting backgrounds or introducing artifacts. One way to assess this is to ask listeners to choose their preferred DE boost levels for both guided and unguided DE in non-simultaneous trials, then compare the boost levels chosen [2]. However, such a method does not allow for direct comparison of ideal dialog boosted and estimated dialog boosted signals. It also requires machine estimation (in this case, using the BSS Eval Toolbox [16]) to characterize the amount of dialog boost achieved by the unguided DE system, but such estimation is itself not necessarily perceptually accurate [17].

We aim to assess DBQ by having listeners directly compare an estimated boosted signal against an ideal

boosted signal, and by explicit characterization of the qualities of the estimated boosted signal. To this end, we conducted two listening tests. We include test items with a range of DNRs (dialog-to-nondialog ratios), dialog mixing types, and background types; which pose a range of difficulty to the SLF+C system; and which have not been used for development of SLF or classification.

6.2 MUSHRA Test with Clean Dialog

For the first of two listening tests, we will require audio for which both the original mix and the clean dialog are available, which allows comparison with a perfectly boosted dialog reference signal. The items used for this test are drawn from a random set of audio clips recorded from a San Francisco Bay Area cable feed of broadcast television in 2018 and 2019. The audio was cached by an automated system which received signals after they had been encoded in AAC format at various bit rates then decoded. Using such signals simulates end-consumer applications where the DE system exists at some post-encode point (e.g. mezzanine or emission) in the content delivery chain.

Name	Gender	Music	Effects	Genre	$\Theta_i/(\pi/2)$	DNR	Median MUSHRA
Test 13	F	X	C	L	0.5	-0.1	69.5
Test 14	M	X	C	L	0.5	2.0	67.0
Test 16	B	X	C	L	0.5	12.8	75.0
Test 26	M	X	R	M	0.1	14.5	87.0
Test 27	M	X	R	M	0.5	15.2	78.0
Test 41	F	A	X	T	0.5	5.5	84.0
Test 43	M	S	X	P	0.2	2.8	79.5
Test 52	M	L	X	D	0.5	13.0	68.5
Test 53	F	X	N	N	0.4	7.3	81.5
Test 54	F	X	C	N	0.5	4.6	89.0
Test 55	M	L	X	P	0.5	3.3	72.0
Test 66	F	S	X	P	0.4	7.2	89.0
Test 83	B	L	C	D	0.5	5.3	67.5
Test 102	B	X	M	L	0.5	-1.4	48.0
Test 107	F	L	X	T	0.5	1.1	72.0

Table 2. Characteristics of MUSHRA test items.

The audio was originally in 5.1 format; an automated system with human verification identified 113 clips of various lengths (each 10 to 15s, with no more than brief portions lacking backgrounds) where there was exclusively dialog in the center channel and exclusively non-dialog in the other channels. To create a clean dialog signal, the center channel was upmixed to stereo; to obtain a clean background signal, the non-center channels were downmixed to stereo. To generate a 9 dB perfect boost version, the clean dialog was added back to the mix with a gain of 10^(9/20)-1.

MUSHRA data set. Of the 113 clips in the original data set, 15 items were quasi-randomly selected to allow representation of various genres, background types, DNRs and speakers' gender presentation. For four of the clips (from genres represented by more than one clip each), the dialog mixing was modified to use a quasi-random non-center panning coefficient. This tests versatility of the DE system because typical content includes non-center-panned dialog. One

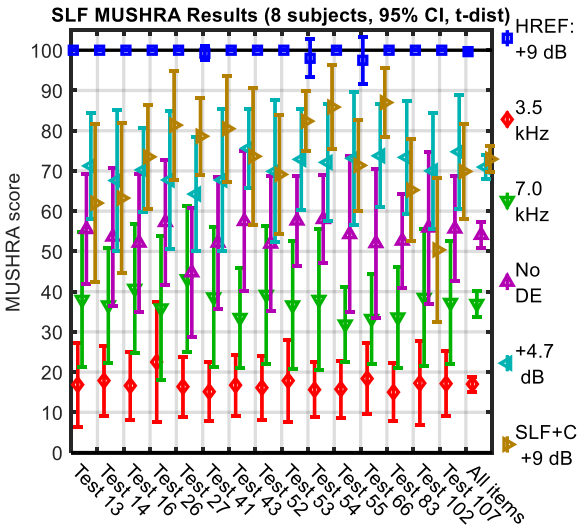

Fig. 9. MUSHRA test results.

pathologically mixed item (Test 102) was deliberately chosen due to complete spatial overlap of loud backgrounds with most of the dialog. A list of abbreviations is shown in Table 1, and characteristics of the items for the MUSHRA test are shown in Table 2, along with summary results from human listening described below.

MUSHRA tests evaluate how well a system under test produces audio which matches a reference. We performed a MUSHRA test following the ITU-R BS.1534-3 specification [18], using eight expert listeners and the MUSHRA test content items described above. The references were the "ideal" or "perfect" 9 dB dialog-boosted signals, and the system under test was the SLF+C system at 9 dB boost. In addition to the two standard anchors (reference signals lowpass filtered at 3500 and 7000 Hz), we also included the original mixes (with no dialog boost) and a 4.7 dB perfect dialog-boosted signal. The latter was motivated by the automated SIR metric (from [19], an update to [16]) which found that the median dialog-to-background ratio increase achieved by the SLF+C system was 4.7 dB. (The same metric found the perfect system's increase was 9.0 dB as expected.)

Results are shown in Fig. 9. Some SLF+C items performed very well (all three erroneous scores of 100 were from different subjects and for the SLF+C system), and overall the system performed similarly to the 4.7 dB perfect boost system. Performance was generally better on higher DNR items but did not vary substantially over other characteristics. Item “Test 102” performed poorly, as expected. See the Conclusion section for further comments.

6.3 Comparison Ratings Test: Mix Only

Using test content for which only the mix is available allows evaluation of arbitrary authentic mixes of interest, even if clean dialog tracks are unavailable. Items of 10-15 sec duration from various sources were selected to allow or increase

No.	Name	Gender	Music	Effects	Genre	Dialog Mixing	Approx. DNR
1	Antiques	F	A	X	T	Moderately reverberant	Moderate-high
2	MiniseriesDrama	B	A	C	T	Reverberant, panned	Low-moderate
3	SoccerBkgDlg	M	X	C	L	Non-center-panned	Moderate
4	MovieDemo5*	F	L	X	T	Highly reverberant	Moderate
5	AutoRacing1	M	X	M,R	M	Center-panned	Very low
6	FamilyDrama	B	A	X	T	Various	Moderate
7	CrimeDrama	B	A	S	T	Various panning	Moderate
8	ImmersiveTrailer***	M	L	M,S	T	Various panning	Low
9	SciFiHorrorMovie	B	S,L	X	T	Highly reverberant, panned	Moderate
10	SoccerGoal	M	X	C	L	Center-panned	Moderate, low
11	SciFiDramaMovie	M	X	M	T	Moderately reverberant	Moderate
12	WesternMovie**	B	A	N	T	Various panning	Low-moderate
Dialog from: *mystical witch character, **boy and girl characters, ***boy and man characters.							

Table 3. Characteristics of mix-only content items.

representation of reverberant dialog mixing, speaking character types, the scripted TV / movie genre, and spatial object backgrounds. Items and characteristics are shown in Table 3. We used these items to assess how well the SLF+C system achieved boost and artifact goals for two boost levels of 9 and 15 dB. We employed a forced-choice comparison test which presents a listener with a randomly ordered pair of items, any two of: (1) unboosted (0 dB) mix signal (2) SLF+C system with 9 dB dialog boost (3) SLF+C with 15 dB dialog boost. The listener is asked: "Given the goals of (1) significantly louder dialog than background (2) minimal artifacts / distortions, which item do you prefer overall? By how much?" The listener specifies item A or B, and a numeric strength rating (minimum increments of tenths of a point) from: (1) not at all (2) slightly (3) moderately (4) very much (5) extremely. We ran seven subjects, all expert listeners with normal hearing.[1]

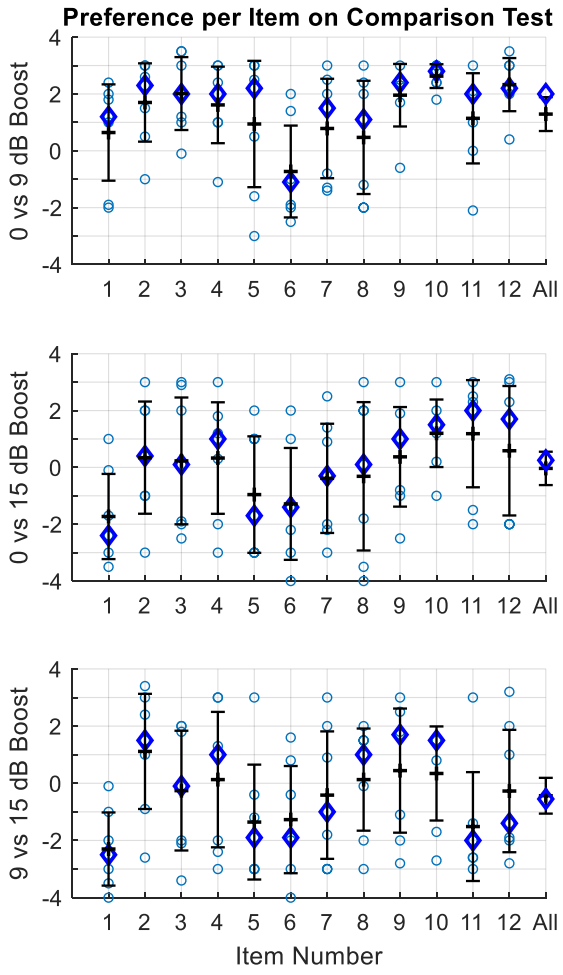

Fig. 10. Comparison test results.

To plot results in a compact fashion, we normalized and combined ratings for each system pair by subtracting 1 from all strength scores, and making the values negative for the first system under test, and positive for the second. For example, in the 0 vs 9 dB comparison, an individual data point preference value of +2 indicates "moderately" preferring the 9 dB boosted signal, and -1 indicates "slightly" preferring the 0 dB signal. Results are shown in Fig. 10 for 0 vs 9 dB, 0 vs 15 dB and 9 vs 15 dB SLF+C boost. The mean and the 95% confidence interval are shown with horizontal marks on a vertical bar for each item and over all items. Medians are shown as blue diamonds, and individual data points per item are shown as small open circles. We see that the 9 dB boosted signal was preferred over no boost in almost all cases, while results were mixed for the other two comparisons. For those comparisons, additional data collected during testing showed that preferences were *not* strongly correlated with the perceived relative dialog level or artifacts in the tested signals (data not shown). Listener comments and informal follow-up testing suggested preferences were often based on the perceived *need* (or lack of need) for dialog boost, rather than *quality* of dialog boost; there was also aversion to excessive overall volume for the 15 dB boosted signals. Future tests will update procedures (see below) to yield better insights into DBQ at high boost levels.

6.4 Discussion

Above, we presented data from human listening tests. Updated or reinterpreted versions of automated

[1]The COVID pandemic interfered with typical procedure for obtaining a larger number of listeners.

source separation metrics also show some potential which we intend to pursue in future work.

We could also improve the listening tests. The MUSHRA evaluation could be updated to allow forgiveness for mutual changes in dialog and background levels (which may be immaterial and mitigated in practice by volume control) by doing a separate pre-exercise which matches a tested item's dialog or background level to that of the reference. A similar exercise (for MUSHRA and comparison test items) could measure boosted signal dialog and background levels to estimate perceived final DNR achieved. This could be used to judge the system (e.g. "fell 2 dB short of boost target") or adapt its output to attempt greater boost at the risk of more artifacts, which could be separately evaluated. Dynamic range issues in testing could be mitigated by allowing a combination of background ducking and dialog boost rather than using boost-only DE.

The SLF system exploits typical differences between dialog and backgrounds with regards to their spatial characteristics, though we did not spatially characterize backgrounds in the evaluation here; future evaluation could do so via a combination of machine and human methods. More generally, for more stringent future testing, we should incorporate ITU-R BS.1116 listening tests, evaluate items with higher levels of dialog boost, or both.

7 Conclusion and Future Work

Above, we described the SLF system, including the SLF model, spatial motivation, system operation, and combination with classification. We described and conducted evaluation which indicated favorable performance. We observed that the system requires little lookahead, memory, computation and training data, which makes it suitable for various DE applications. We now consider areas for future work.

The SLF+C system described here is designed to accept stereo content as input. The system has been repurposed to accept input of Left, Center and Right channels in 5.1 or higher channel count content, where dialog is commonly mixed; this will be described in a future publication. (For object-based audio with objects of unknown content, the source separation problem is substantially different, becoming one primarily of dialog detection within objects.) For cases with mono input, the SLF system can accept a stereo input formed from duplicates of the mono channel, though this leads to trivial θ and ϕ data; performance decreases as the system must act only based on U.

Generally, inputs with trivial (e.g. constant scalar) data for θ and ϕ, are challenging, as are cases where backgrounds are mixed identically to time-coincident dialog. In such cases, pairing SLF+C with a technology built to extract mono dialog has shown promise; this technique is also effective if SLF+C performs moderately well. When SLF is strained by spatially concentrated backgrounds, applying classification to each of the SLF output and residual could mitigate; we will explore this. Currently the data in S&S parameters is not exploited except for extraction, although it may indicate spatially concentrated interferers, multiple target sources, or strain. Work is ongoing to exploit S&S data, and to update its interpretation when the target signal is not dialog, e.g. for music-only applications.

We noted above that the adaptive S&S extraction approach allows extraction of a quasi-infinite number of spatially identifiable sources from a single trained filter. However, this does not guarantee that the extraction quality will match what individually trained filters would have achieved, as the adaptation is based on limited training data. Such training could be significantly expanded. The 1.5x squeeze limit could be made to vary based on other detected features. We look forward to pursuing all the above ideas.

References

[1] A. Master and H. Muesch, "A Model to Predict the Impact of Dialog Enhancement or Mix Ratio on a Large Audience," in *AES 149th Convention*, New York, 2020.

[2] M. Torcoli, J. Herre, H. Fuchs, J. Paulus and C. Uhle, “The Adjustment/Satisfaction Test (A/ST) for the Evaluation of Personalization in Broadcast Services and Its Application to Dialogue Enhancement,” *IEEE Transactions on Broadcasting*, vol. 64, pp. 524-538, June 2018.

[3] E. Vincent, T. Virtanen and S. Gannot, Audio Source Separation and Speech Enhancement, Wiley, 2018.

[4] J.-T. Chien, Source Separation and Machine Learning, Academic Press, 2018.

[5] A. Master, "Stereo Music Source Separation via Bayesian Modeling," Ph.D. Dissertation, Dept of Electrical Engineering, Stanford University, 2006.

[6] M. Michelashvili and L. Wolf, “Speech Denoising by Accumulating Per-Frequency Modeling Fluctuations,” arXiv.org, 2019.

[7] T.-C. Zorilă, Y. Stylianou, S. Flanagan and B. Moore, “Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss,” *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 189-196, 2017.

[8] J. T. Geiger, P. Grosche and Y. L. Parodi, “Dialogue Enhancement of Stereo Sound,” in *European Signal Processing Conference*, Nice, 2015.

[9] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830-1847, 2004.

[10] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, “Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149-157, 2001.

[11] S. Araki, S. Makino, H. Sawada and R. Mukai, “Reducing Music Noise by a Fine-Shift Overlap-Add Method Applied to Source Separation Using a Time-Frequency Mask,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, 2005.

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

[13] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff and J. Hershey, “The Phasebook: Building Complex Masks via Discrete Representations for Source Separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019.

[14] L. Lu, H.-J. Zhang and S. Li, “Content-based Audio Classification and Segmentation by Using Support Vector Machines,” *ACM Multimedia Systems Journal* , vol. 8, no. 6, pp. 482-492, 2003.

[15] B. Cheng and L. Lu, “Audio classification method and system”. US Patent US8892231B2, 18 11 2014.

[16] E. Vincent, Gribonval, R and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.

[17] J. Le Roux, H. Erdogan, J. R. Herskey and S. Wisdom, “SDR – Half-baked or Well Done?,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019.

[18] ITU-R, “Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems,” International Telecommunications Union, Geneva, 2015.

[19] V. Emiya, E. Vincent, N. Harlander and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Transactions on Audio Speech, and Language Processing*, vol. 19, no. 7, pp. 2046-2057, 2011.