OPEN ACCESS Freely available online

A. S. Master, L. Lu and N. Swedlow, "Stereo Speech Enhancement
Using Custom Mid-Side Signals and Monaural Processing"
J. Audio Eng. Soc., vol. 71, no. 7/8, pp. 431–440, (2023 July/August).
DOI: https://doi.org/10.17743/jaes.2022.0070

Stereo Speech Enhancement Using Custom Mid-Side Signals and Monaural Processing

AARON S. MASTER, LIE LU, AND NATHAN SWEDLOW
(amaster@gmail.com) (Lie.Lu@dolby.com) (Nathan.Swedlow@dolby.com)

Dolby Laboratories, Inc., San Francisco, CA

Speech enhancement (SE) systems typically operate on monaural input and are used for applications including voice communications and capture cleanup for user-generated content. Recent advancements and changes in the devices used for these applications are likely to lead to an increase in the amount of two-channel content for the same applications. However, SE systems are typically designed for monaural input; stereo results produced using trivial methods such as channel-independent or mid-side processing may be unsatisfactory, including substantial speech distortions. To address this, the authors propose a system that creates a novel representation of stereo signals called custom mid-side signals (CMSS). CMSS allow benefits of mid-side signals for center-panned speech to be extended to a much larger class of input signals. This, in turn, allows any existing monaural SE system to operate as an efficient stereo system by processing the custom mid signal. This paper describes how the parameters needed for CMSS can be efficiently estimated by a component of the spatio-level–filtering source separation system. Subjective listening using state-of-the-art deep learning–based SE systems on stereo content with various speech mixing styles shows that CMSS processing leads to improved speech quality at approximately half the cost of channel-independent processing.

0 INTRODUCTION

Speech enhancement (SE) is a technology that aims to reduce or eliminate background noise while preserving speech quality, typically for monaural audio signals in applications including voice communications and capture clean-up. State-of-the-art deep learning–based SE systems include FullSubnet [1], Dual-Path Convolution Recurrent Network (DPCRN) [2], and NSNet2 [3], which is the official baseline for the Deep Noise Suppression challenge [4]. These technologies are relatively mature for monaural input but are not designed for stereo (or higher channel count) inputs; pilot testing of these systems, using channel-independent processing of stereo signals with various styles of speech mixing, found that they tend to produce unsatisfactory outputs with significant noise leakage and speech distortion. For channel-independent processing, the SE systems are likely to exhibit unequal strain vs. time and frequency in each channel, and the unequal imperfections may be perceived as additional distortions.

Due to recent trends in hardware and operating systems, there is likely to be a large increase in the amount of stereo data available for SE applications. One major consumer electronics company recently updated their operating system to allow access to both mics on their mobile phones and tablets [5]; the company has over 1.8 billion active devices [6]. An operating system provider has facilitated the inclusion of microphone arrays (including two-mic arrays) on laptops for several years [7]; it is possible they will also facilitate accessing array signals directly. Other advancements include binaural microphones on headphones (e.g., [8, 9]) and an increase in the availability of affordable peripheral stereo microphones.

This likely increase in the availability of stereo inputs to SE systems that are not designed for such inputs presents an opportunity for improvement. This paper presents a system that allows an existing monaural SE system to process a new type of signal based on a spatially dynamic version of mid-side signals. To develop an understanding of the new signal, the authors will describe the special relationship between center-panned speech target sources and standard mid-side signals, namely that the mid signal boosts the target source relative to other signals, whereas the side signal suppresses it.

The authors will then describe how this concept may be generalized to target sources that are *not* center-panned, whose mixing may be estimated using a technique described in [10]. This allows creation of custom mid-side signals (CMSS) that allow arbitrarily mixed speech to receive the same benefit as center-panned speech in standard mid-side signals. An SE system may process the custom mid signal (CMS), instead of processing standard mid, side, or channel signals.

Using a subjective listening test, the authors demonstrate that for state-of-the-art deep-learning based SE systems, this approach allows for significantly improved speech output quality at approximately half the processing cost of channel-independent processing.

This paper is organized as follows. Sec. 1 reviews mid-side signals and describes how they can be generalized and customized using detectable mixing parameters to create CMSS. Sec. 2 describes how to use CMSS for efficient, high-quality processing of stereo input by a monaural SE system. Sec. 3 presents results from a subjective listening test of SE systems, which compares the perceived qualities of the proposed method with those of a channel-independent baseline. Sec. 4 concludes the paper with a summary and discussion of future work.

1 MID-SIDE SIGNALS

This section describes how CMSS are created and motivate their use. To do so, the first subsection begins by reviewing standard mid-side signals and notes their special relationship with center-panned sources. The second subsection will describe a more general target source mixing model, which uses additional parameters Θ_1 and Φ_1 to describe mixing of sources that are not necessarily center-panned. The third subsection describes how to create generalized mid-side signals that provide mid-side benefits to sources that are mixed using the more general model. The fourth subsection describes how the concept of generalized mid-side signals may be effectively implemented for real-world signals by dynamically estimating Θ_1 and Φ_1 over time and frequency using a technique in [10]. The resulting signals are termed CMSS.

1.1 Standard Mid-Side

The authors presently summarize the standard mid-side signal decomposition (see, e.g., [11]). Mid-side microphone capture is not presently described, although the authors note that a mid-side signal decomposition of stereo signals may approximately recover the components signals in a mid-side recording; see, e.g., [12]. The standard mid-side decomposition of a stereo signal is as follows:

$$\begin{aligned} M &= 0.5\,(L + R) \\ S &= 0.5\,(L - R)' \end{aligned} \tag{1}$$

where M is the mid signal, S is the side signal, L is the left channel signal, and R is the right channel signal. Because the operations used here are linear, this calculation may occur in the time domain or a time-frequency domain such as the short-time Fourier transform (STFT) domain. Below, the authors will work in the STFT domain. The original stereo channels may be recovered via

$$\begin{aligned} L &= M + S \\ R &= M - S. \end{aligned} \tag{2}$$

For inputs that contain a center-panned target signal of interest, the mid signal will contain the target signal (and likely, other sounds), whereas the side signal will be devoid of the target signal but will contain other non–center-panned sounds, depending on how they are mixed to the two channels. (See, e.g., sec. 4.4 of [13].) For stereo signals with a center-panned target signal, the mid-side representation may be thought of as providing mild source separation or source boosting. The mid signal will increase the relative level of center-panned in-phase signal components (and of components that are approximately so) while attenuating others; the side signal will completely attenuate center-panned signal components and will boost out-of-phase components.

This special relationship between a center-panned target source and the mid-side signals can be beneficial to a processing system that seeks to enhance a target source, as will be described below. The subsections that follow will describe how standard mid-side signals can be generalized and customized to extend this benefit to a much larger class of stereo input signals. In order to do so, the authors next introduce a more general source mixing model that describes target sources that are not necessarily center-panned.

1.2 Generalized Mixing Model

In order to develop a more generalized version of mid-side signals, a more generalized mixing model for the target source must be developed. This subsection describes such a model; the next subsection will then develop a generalized version of mid-side signals based on this model.

This model considers how a monaural source S_1 with magnitude $|S_1|$ and phase Ψ_1 for each STFT tile $(\omega,\ t)$ is mixed to two channels (L and R) in STFT space. First, the mono source is defined as

$$S_1\,(\omega, t) = |S_1\,(\omega,\ t)|\exp\,(i\ \Psi_1\,(\omega,\ t))\,, \tag{3}$$

where it is noted that values of S_1 exist for each STFT tile of frequency bin ω and frame t; going forward, $S_1(\omega, t)$ (and similar such quantities) shall be abbreviated as S_1 (and similar) for simplicity.

The mixing shall be modeled with regard to inter-channel level difference (ILD) and inter-channel phase difference (IPD). For purposes of developing generalized mid-side equations, each of these quantities is temporarily treated as a single fixed value for all times and frequencies.

In practice, these quantities will vary; the IPD must be allowed to vary vs. frequency and time in order to model sources mixed with reverberation or inter-channel delay. IPD concentrations can also be easier to estimate than delay (see Sec. 3 of [10]), especially in cases in which microphone geometry and movement is unknown or difficult to model, such as for binaural capture. The ILD is modeled by a panning coefficient Θ_1 ranging from 0 (pure left) to $\pi/2$ (pure right) under the constant power panning law [14], leading to the following values of S_1 in the L and R channels when the IPD is zero:

$$\begin{aligned} L &= S_1 \cos(\Theta_1) \\ R &= S_1 \sin(\Theta_1)\,. \end{aligned} \tag{4}$$

The IPD is described by the parameter Φ_1. When mixing is modeled with nonzero IPD, the relationship between Ψ_1 and Φ_1 must be explicitly defined. One option [15] is to declare the left channel phase to be the true source phase, in which case the phase difference applies only to the right channel. However, doing this creates a problem when describing source phase and IPD for an extreme right panned source. In such cases, the left channel's phase information is unrelated to the target source and effectively random, influenced by values in the noise floor or backgrounds. Modeling the IPD as split evenly between the channels creates a similar problem; the left channel is still random.

To address this, the authors use a mixing model in which the channel where the source is stronger in power has proportionally closer phase to the source phase Ψ_1 and the other channel is proportionally less close to Ψ_1, and dictated by Φ_1. By careful selection of these proportions based on Θ_1 data, it can also be ensured that the phase difference between the channels still equals Φ_1. The L and R channels are thus modeled as

$$\begin{aligned} L &= S_1 \cos(\Theta_1) \exp(i\ \Phi_1\ \sin^2\Theta_1) \\ &= |S_1|\ \exp(i\ \Psi_1) \cos\Theta_1 \exp(i\ \Phi_1 \sin^2\Theta_1) \\ &= |S_1| \cos\Theta_1 \exp(i\ (\Psi_1 + \Phi_1 \sin^2\Theta_1)) \\ R &= S_1 \sin(\Theta_1) \exp(-i \Phi_1\ \cos^2\Theta_1) \\ &= |S_1| \exp(i\ \Psi_1)\ \sin(\Theta_1) \exp(-i\ \Phi_1 \cos^2\Theta_1) \\ &= |S_1| \sin(\Theta_1) \exp(i\ (\Psi_1 - \Phi_1 \cos^2\Theta_1)). \end{aligned} \tag{5}$$

It can be seen from examining Eq. (5) that, for tiles in which only the target source is present, the IPD, calculated via $\angle(L/R)$ (see [10]) will equal Φ_1. Similarly, the ILD, calculated via $\arctan(R/L)$ will equal Θ_1. APPENDIX A further explores these calculations and their relationship with $|S_1|$ and Ψ_1.

1.3 Generalized Mid-Side

Now that a generalized target source mixing model has been described, generalized mid-side signals may be described based on this model. It can be seen that for a target source mixed as specified in the Eq. (5) above, the authors can define generalized mid and side signals that will have the boosting and elimination properties that standard mid-side signals have for center-panned signals. Eq. (6) below specifies such signals, which are termed *generalized mid-side signals.* (For a more detailed derivation of the side signal, see “normalized weighted subtraction” on p. 38 of [13]; the derivation of the mid signal is similar.)

$$\begin{aligned} M &= c_1 L + c_2 R \\ S &= c_3 L + c_4 R \end{aligned} \tag{6}$$

where

$$\begin{aligned} c_1 &= \cos\Theta_1 \exp(-i \Phi_1 \sin^2\Theta_1) \\ c_2 &= \sin\Theta_1\ \exp(i \Phi_1 \cos^2\Theta_1) \\ c_3 &= \sin\Theta_1\ \exp(-i \Phi_1 \sin^2\Theta_1) \\ c_4 &= \cos\Theta_1\ \exp(i \Phi_1 \cos^2\Theta_1). \end{aligned}$$

The *inversion* equations (or *stereo reconstruction equations*), which return to conventional stereo signals, are

$$\begin{aligned} L &= (M \cos\Theta_1 + S \sin\Theta_1) \exp(i \Phi_1 \sin^2\Theta_1) \\ R &= (M \sin\Theta_1 -\ S \cos\Theta_1) \exp(-i \Phi_1 \cos^2\Theta_1). \end{aligned} \tag{7}$$

As a point of clarification, the authors note that, for a center-panned target source (for which $\Theta_1 = \pi/4$ and $\Phi_1 = 0$), these generalized mid-side equations (and thus the inversion equations) use different scaling than the standard mid-side equations but are otherwise identical. The inversion equations still recover L and R channels with the original scale intact. Using the noted values of Θ_1 and Φ_1 leads to

$$\begin{aligned} M &= \sqrt{2}/2\ (L + R) \\ S &= \sqrt{2}/2\ (L - R) \\ L &= \sqrt{2}/2\ (M + S) \\ R &= \sqrt{2}/2\ (M -\ S)\,. \end{aligned} \tag{8}$$

1.4 Custom Mid-Side

The generalized mid-side signal concept of the previous subsection can be further expanded by allowing the parameters Θ_1 and Φ_1 to vary vs. time and frequency. This effectively allows for special mid and side signals that track target source signals whose mixing parameters are dynamic. An example is a target source that moves (relative to stereo capture microphones) whose corresponding Θ_1 and Φ_1 values will vary with time. (And as noted above, Φ_1 must be allowed to vary with time and frequency to model reverberant sources or those mixed with inter-channel delay.) The time-varying and frequency-varying mid-side signals are termed CMSS. Their mathematical expressions are identical to those introduced in the previous subsection, with additional flexibility in that Θ_1 and Φ_1 are allowed to vary by frequency sub-band b and time frame t; they become $\Theta_1(b, t)$ and $\Phi_1(b, t)$.

In order to generate CMSS, $\Theta_1(b, t)$ and $\Phi_1(b, t)$ which are termed *dynamic mixing parameters*, must be estimated and applied. To estimate these parameters, the authors use a process described in [10] termed *spatially identifiable subband source detection* (SISSD) (and where the parameters are also termed *thetaMiddle* and *phiMiddle*). The parameters are estimated at a frequency granularity of one parameter per quasi-octave frequency sub-band (sub-band edges are [0; 400; 800; 1,600; 3,200; 6,400; 13,200; 24,000] Hz) and updated once every 1,024 samples for 48-kHz sampled audio.

The choice of granularity is critical; if very large frequency bands are chosen, reverberant sources cannot be well estimated, and if one chooses to update values infrequently vs. time. Then rapidly moving sources cannot be tracked. However, choosing too fine a granularity leads to unreliable and unstable estimates. In the most extreme case, one could estimate Θ_1 and Φ_1 for each STFT tile that leads to the mid signal containing all energy and the side signal containing none. In that case, the parameters do not characterize meaningful target source mixing but rather the individual statistics of a single tile or micro-region. Granularity trade-offs are further explored in Secs. 2 and 3 of [10].

Using the noted update rate and frequency bands was found to be perceptually effective when estimating dialog for dialog enhancement applications for typical challenging entertainment content mixes [10] and in pilot tests of pathological mixes with multiple simultaneous human speakers at different spatial locations. In such cases, the parameters update often enough vs. time and frequency that they can alternate between tracked human speakers and still capture most of the speakers' perceptually salient energy. This is in general agreement with related speech source separation techniques (e.g., [15]) that rely on the relatively infrequent overlap of speech energy from multiple independent speakers in STFT space. For cases in which the SISSD detects both target speech and non-speech sounds, additional processing (such as the SE systems described below) serves to eliminate or attenuate such non-speech sounds.

The authors also note that the model used here characterizes the mixing of a monaural source to two channels using only the two (time-varying and frequency-varying) parameters $\Theta_1(b, t)$ and $\Phi_1(b, t)$. In practice, some sources, especially those mixed with heavy reverberation, cannot be so simply characterized. In such instances, the model used here effectively becomes a single wavefront approximation of the source mixing (see e.g., APPENDIX A of [13]), which is found to still lead to a large quality improvement over existing baselines.

2 CMSS AND SE

2.1 Benefits of CMSS to SE

To summarize the above discussion, the authors have now greatly expanded the types of signals that can have a special relationship with a mid-side representation. A center-panned target source is boosted in a standard mid signal and eliminated from a standard side signal. With CMSS, however, any spatially concentrated target source detected by SISSD is boosted in the CMS and eliminated or attenuated in the custom side signal.

The authors now revisit why this is beneficial. Knowing that a mid signal will contain the target source, whereas the side signal will suppress it, means that an SE system could process only the mid signal and still capture the target source. However, for standard mid-side signals, this property only holds if the target source is actually center-panned. If the speech is captured or mixed with inter-channel delay or is at a higher level in one channel than the other, and the standard mid signal is used regardless, the mid signal may have a lower signal-to-noise ratio (SNR) than the side signal or channel signals. If an SE system were to process only the standard mid signal for such a case, it would lead to a result in which speech was underestimated or even entirely missed. By using SISSD to calculate parameters for CMSS, one can obtain a special, robust mid signal, which boosts a target source, while a special side signal eliminates or attenuates it. An SE system can then process only the CMS before reconstructing the output.

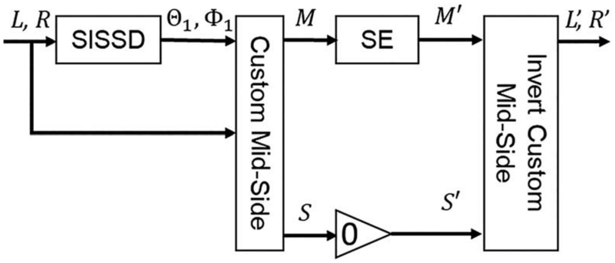

Fig. 1. Stereo Speech Enhancement using CMSS.

2.2 Proposed Signal Flow and SE

Given the above description of how CMSS are obtained, the authors now describe a stereo processing method, depicted in Fig. 1, which can utilize any monaural SE system. It is observed that the input signal enters as a stereo pair (L, R), which the SISSD processes to obtain $\Theta_1(b, t)$ and $\Phi_1(b, t)$, which, along with the stereo input signal, are passed to the CMSS generator. The generator produces the CMSS as described by Eq. (6) above. Note that CMS does not include the side signal. The CMS is passed to a monaural SE system, whose output is termed M'. The side signal is set to zero, and this signal is called S'. Finally, M' and S' are passed to a module that calculates L' and R' according to the CMSS reconstruction Eq. (7), thereby producing the system output. This output shall be termed the CMS processed version; these signals are the ones evaluated as the CMS condition in the evaluation section.

In principle, any monaural SE system could be used in the overall system design proposed. For evaluation, two such systems are used, one developed internally, and the other an available state-of-the-art system. The internal system, which is termed U-NetFB, is an SE network with a U-Net type architecture, similar in concept to [16–18], but the inputs are frequency band energies, rather than STFT bin values, and the outputs are real-valued frequency band softmask values. The second system, DPCRN [2], is chosen based on its relatively high performance compared with other state-of-the-art systems in a separate pilot test. Each of these systems is deep learning–based and is of much greater computational expense than the SISSD used to estimate $\Theta_1(b, t)$ and $\Phi_1(b, t)$. As a result, the computational cost of the proposed system is similar to a single instance of either SE system without the SISSD; adding a second instance of an SE system, however, approximately doubles the cost.

2.3 Alternative Processing Options

An overall system in which the SE component processes CMS only has been proposed. However, there are alternatives that are now considered, along with their potential benefits and drawbacks. One such method, channel-independent processing, will be proposed as a baseline for comparison with CMS processing. There will also be descriptions of why other options were not used as a proposed system or baseline.

2.3.1 CMSS

First, the system proposed in the previous subsection is considered, but with the custom side signal also processed by a second instance of SE rather than set to zero. This option was initially considered because it is more spatially exhaustive than the proposed method: processing both custom mid and side signals ensures that if speech is present, some version of it will be processed. However, testing of various real-world signals (including stereo-captured content and professionally generated stereo content) found that it is relatively rare for the custom side signal to contain significant undistorted speech energy compared with the CMS. Processing a signal that contains little to no actual speech risks that speech will be erroneously detected, leading to perceptible errors. As noted above, an extra instance of SE also approximately doubles the computational cost. Given the cost and risks, this option was declined for now. Nonetheless, the authors consider this an area for future work because there are likely to be some signals, namely those for which the SISSD performs imperfectly, that benefit from CMSS processing vs. CMS-only processing.

2.3.2 Standard Mid-Only

Another variation on the proposed system is to process only the mid signal, but for a standard mid-side decomposition. For center-panned speech signals, this will have similar performance as CMS processing, whereas for other types of mixing, speech will be attenuated or missed entirely as noted above. For this reason, this option was not considered as a viable alternative or meaningful baseline.

2.3.3 Standard Mid-Side

A related idea is to process both standard mid and side signals for a given input, because they are similarly spatially exhaustive. For center-panned speech mixing, this approach has similar risks and costs (two SE instances) compared with CMSS processing, but for other speech mixing, this approach is similar in performance to channel-independent processing, because the standard mid-side signal pair may be understood as a spatial rotation of the original stereo signal pairs (see, e.g., Sec. 4.4 of [13]). Given the costs and risks, the authors declined this option.

2.3.4 Alternative Center

Another novel option is to create an *alternative center* stereo signal from the CMSS by using standard inversion Eqs. (2) or (8), which do not consider $\Theta_1(b, t)$ and $\Phi_1(b, t)$, rather than the CMSS inversions in Eq. (7), which do. In this case, the reconstructed stereo signal will re-mix the spatial concentrations detected by the SISSD to make them center-panned. Doing so allows for the processing of the alternative center signal using any center-panned or center-biased system, including those designed for enhancing dialog in entertainment content such as [19], which targets center-panned speech, or [20], which favors center sources by using an ILD-based mapping. Because these systems were not available for processing private data, these options were not pursued. (Alternative center signals also allow a simplification of the system proposed in [10], in which the described adaptations for non–center-panned sources become unnecessary.) The authors plan to explore these ideas in future work.

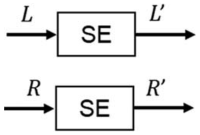

Fig. 2. Channel-Independent SE Processing.

2.3.5 Channel Independent

Perhaps the most obvious alternative processing option is channel-independent processing. In this case, depicted in Fig. 2, each of the stereo input channels is processed separately by an SE system. As noted above, for non–center-panned speech mixing in which speech exists in both channels, such processing is conceptually similar to processing both standard mid and side signals. The significant difference is for center-panned speech mixing, for which channel-independent processing incurs the risk of having different errors in each channel, which may lead to outputs in which listeners perceive distortion. Given that both center-panned and other dialog mixing will be used in the evaluation, channel-independent processing was chosen as a baseline to allow for meaningful comparisons with CMSS in the greatest number of cases. Processing only the standard mid signal described above, for example, would produce essentially identical results as CMS processing for center-panned speech, preventing meaningful comparison.

2.4 Alternative Output Format Options

For the proposed CMS processing system and the baseline channel-independent systems, it can be seen that stereo signals are always output and that the architectures implicitly or explicitly preserve the spatial information of the estimated speech: an input with a center-left speech source should lead to an output with a center-left speech source. This is viewed as the more challenging case; generating a mono signal or trivial stereo signal (e.g., one with both channels identical) does not require the processing system to maintain spatial fidelity. For this reason, stereo output is chosen for use in the evaluation.

However, there may be applications for which spatial fidelity is not required or cannot be included, for example, for voice communication systems that only transmit a mono signal to a listener. For CMS-only processing, a mono output may be obtained by having the system directly output the processed CMS rather than use the inversion Eq. (7). For channel-independent processing, a downmix can be formed. The quality of mono output signals will be considered in future work. For the present, subjects were asked in the evaluation to independently assess speech quality and spatial quality. More details will be described for the evaluation in the next section.

3 EVALUATION

The authors now describe subjective listening tests that were used to compare the performance of the two systems described in the previous section: CMS processing and a channel-independent baseline. The test content, processing, methodology, and results are described.

3.1 Test Content

For input content, stereo items with a variety of dialog mixing styles are used. As noted in the introduction, the authors ultimately aim for the proposed system to process content from a variety of potential stereo sources including two-mic mobile devices, two-mic laptops, binaural headphones, and stereo external mics, all of which have device-dependent types of speech mixing. Pilot tests have been done of the proposed and baseline systems on stereo inputs captured from known and unknown stereo recording devices, as well as on stereo professionally generated content (PGC), i.e., typical TV and movie content, and found results to be broadly similar in nature for low and moderate SNR content. This is attributed to the generality of the SISSD upon which CMSS are based; it is made to detect spatial concentrations of energy, whether they indicate a panned source with some ILD, such as is common in PGC, or a source mixed with inter-channel delay or reverb as is expected for environmental capture.

The present evaluation uses specific PGC items for which the dialog mixing styles can be concisely and accurately described. This allows for evaluation of the proposed and baseline systems for these specific kinds of mixing. Table 1 describes the items by their background type (given as item name), approximate SNR (“Low” indicating less than approximately 5 dB and “Mod” indicating approximately 5–10 dB), speaker gender presentation (in which “B” indicates both male and female speech in the same item), speech mixing style (C indicating center-panned, L-C center-left, and C-R center-right), and genre.

Table 1. Content items for evaluation.

Name	SNR	Gender	Dialog Mixing	Genre
CrowdSing	Low	M	C	Sports
RaceCars	Low	M	C	Motorsports
ShipCrew	Mod	B	C	Scripted
SciFiSFX	Mod	F	C	Movie
Outdoors2	Mod	M	C	Scripted
Bobsled	Low	B	C	Sports
Orchestral	Low	F	C	Ad
UrbanSFX	Mod	M	Varies	Scripted
Cheering	Low	M	C-R	Sports
HallDin	Mod	M	Reverb	Movie
Outdoors	Mod	F	L-C	News
PopMusic	Mod	F	L-C	Ad

Table 2. Attribute descriptions.

Attribute	Description
Preference	Overall preference. Subjects were instructed to identify whether they preferred stimulus A or stimulus B.
Speech Quality	Identification of speech distortion. Subjects identified which signal had *less* distorted speech.
Less Non-Speech	Identification of non-speech sounds. Subjects identified which signal had *less* perceptible non-speech sounds.
Spatial Quality	Naturalness of spatial image. Subjects identified which system sounded more spatially natural and in-line with their expectations for a high-quality experience.

3.2 Test SE Systems

To create the signals used in the test, the processing shown in Figs. 1 and 2 is used. For the SE systems, the U-NetFB system described in Sec. 2 is used for both the CMS and baseline conditions. That is, for CMS, one instance of U-NetFB is run on the CMS only. For the channel-independent baseline, two instances of U-NetFB are run, one on each channel. Additional listening was also done using DPCRN, with similar structure: for the CMS condition, one instance of DPCRN processed the CMS, and for the channel-independent baseline, two instances of DPCRN were used, one for each channel.

3.3 Test Methodology

Nine subjects participated in this experiment and all participants were highly trained in critical evaluation of audio signals. This test was performed in a quiet listening environment with high-quality headphones. Subjects were presented with 12 pairwise comparisons in which they evaluated four attributes per comparison. In each comparison, subjects were presented with two test stimuli (stimulus A and stimulus B)—each stimulus per trial contained identical source content that was prepared using either CMS or the channel-independent baseline (CI). The system ordering was randomized across all 12 trials. Subjects were invited to loop sub-sections of the content and to freely switch back and forth between stimulus A and stimulus B in each trial. The four attributes under test were overall preference, speech quality, less non-speech, and spatial quality, which are defined in Table 2.

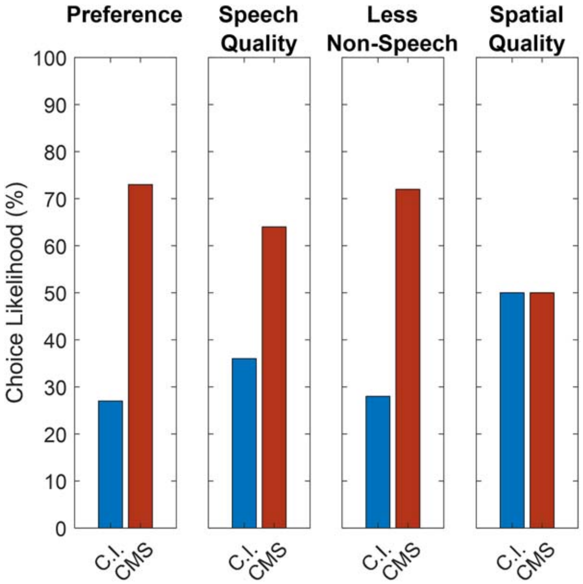

Fig. 3. Choice likelihood across all items and subjects for each question.

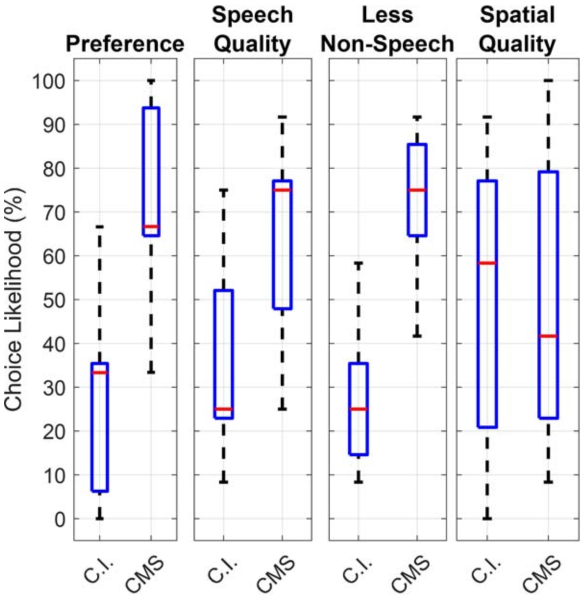

Fig. 4. Box plots showing distributions on choice likelihood across all items and subjects for each question.

3.4 Results

Results are shown in Figs. 3, 4, 5, and 6. Fig. 3 shows that CMS was selected as the preferred system 73% of the time compared to the CI system, across all subjects and all content items. Subjects indicated that CMS processing resulted in better speech quality in 64% of the comparisons under test. Subjects indicated that CMS produced less perceptible non-speech sounds in 72% of the comparisons under evaluation. There was not a notable difference in spatial quality performance across the systems under test. (Additional, less formal listening considering the same content items, conditions and attributes, but with DPCRN [2] instead of U-NetFB for SE processing, found similar audio characteristics.)

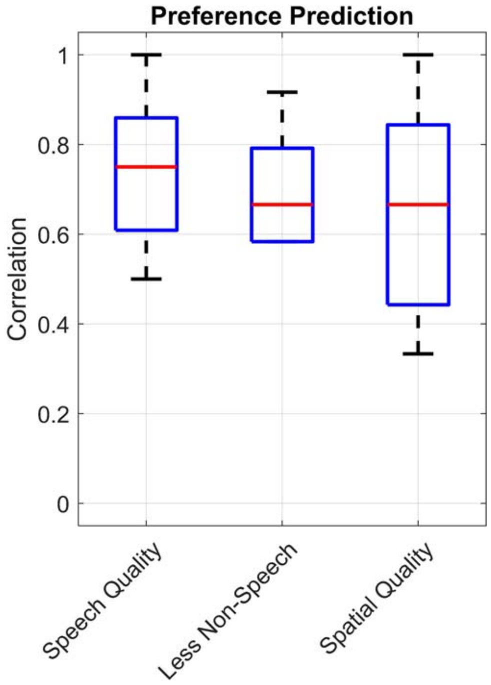

Fig. 5. Correlation between questions and preferences across subjects.

There was substantial variation between subjects across all attributes under evaluation. Fig. 4 depicts the distribution of choice likelihood across all subjects per attribute. This distribution will likely tighten with additional data collection.

Additional analysis found that better speech quality was slightly more correlated to a selection of preference compared to less non-speech and spatial quality (Fig. 5). A considerably larger correlative range for spatial quality is seen relative to both speech quality and less non-speech, indicating that for some subjects, spatial quality did not substantially influence a preference outcome. Figs. 3 and 4 are consistent with this result. There is no overall trade-off observed between preference and processing cost, because the less costly system was also the more preferred system overall.

Fig. 6 shows the choice likelihood for each of the four attributes for each content item listed in Table 1. Additionally, the two rightmost bar pairs for each subplot show results for (1) all items with center dialog mixing and (2) all other items. There is no observed significant difference in choice likelihood for any of the four attributes with regard to mixing type. The single item for which CMSS was less preferred overall, “Bobsled,” includes Bobsled sound effects that are mixed the same as the dialog (center). In future work, the authors will explore how such backgrounds can impact CMSS and CI processing.

In future work, the authors will also consider how to use existing monaural automated metrics for SE, or novel ones, to evaluate CMS and baseline performance. A pilot investigation found that using existing monaural automated SE metrics to evaluate the stereo CMS and baseline results by averaging the metric values for each channel did not yield

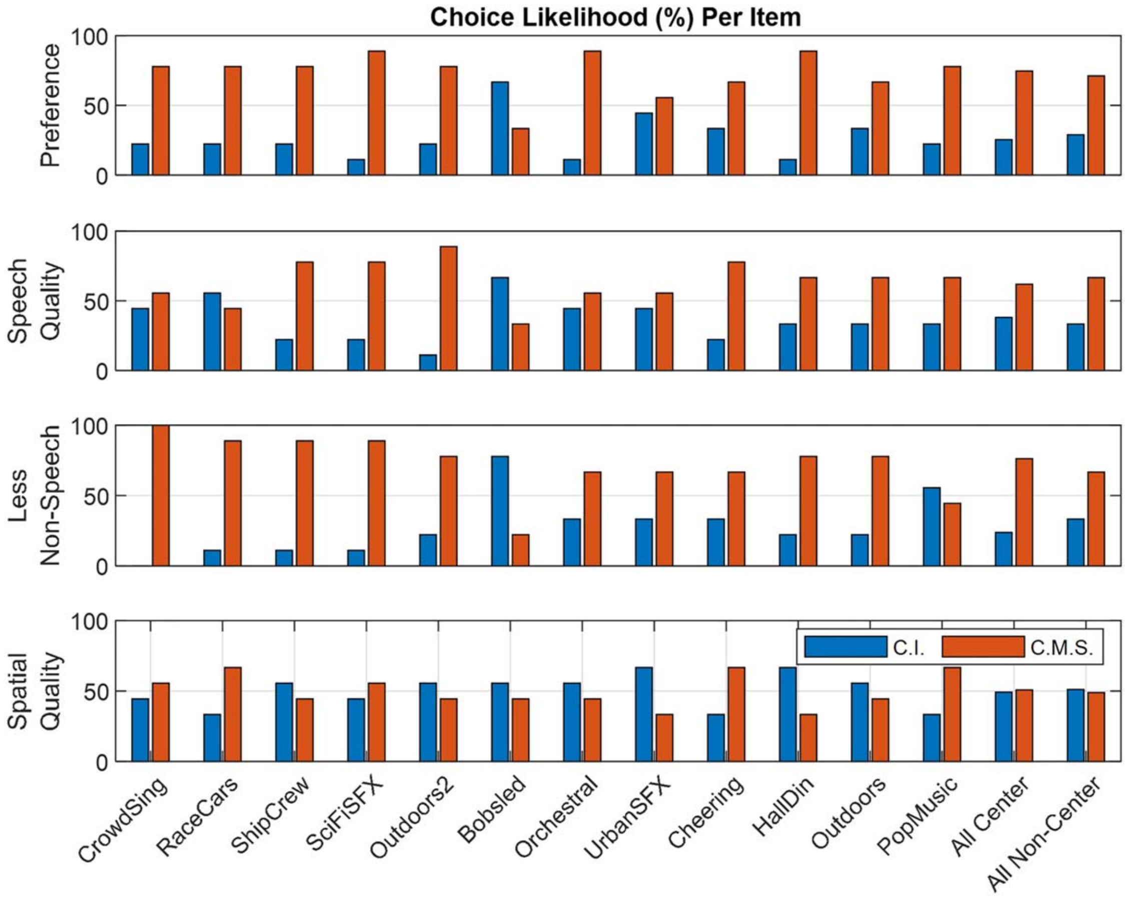

Fig. 6. Choice likelihood for individual content items, all center dialog items, and all non-center dialog items.

results with a meaningful relationship with the subjective data.

4 CONCLUSION AND FUTURE WORK

A theory was developed supporting use of CMS as input to any SE system. The theory was based on the idea that the benefits of mid-side signals for center-panned speech sources could be expanded to a much larger class of signals containing speech by using CMSS informed by the SISSD of [10]. An evaluation on various types of speech mixing provided evidence to support the present theory, because the CMS processed items were preferred over items processed by the same SE system in a channel-independent configuration. This occurred even with the CMS system using only approximately half the computation of the channel-independent system.

In future work, the authors will perform additional testing of content captured in specific real-world contexts, namely via the two mics on specific, commonly used devices, in specific environments, for applications including UGC capture and voice communications. In particular, the authors will attempt to find or generate content that contains spatially concentrated non-speech sounds, which is expected to be able to strain the SISSD components of the proposed system [10]. Early pilot testing of UGC content signals from a variety of sources of capture found results that were broadly similar to those on the tested content items for low and moderate SNRs. For all inputs, it will be investigated whether results can be improved by running SE on both the custom mid and custom side signals.

It was noted that existing automated SE metrics did not provide data with a meaningful relationship with the subjective data presented here. The authors will investigate the underlying causes and consider modifications of these metrics or new ones.

The authors thank Heidi-Maria Lehtonen, Scott Norcross, Jonas Samuelsson, Xiaoyu Liu, Dan Darcy, Audrey Howard, and Libby Purtill for their assistance with this publication.

6 REFERENCES

[1] X. Hao, X. Su, R. Horaud, and X. Li, “Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6633–6637 (Toronto, Canada) (2021 Jun.). https://doi.org/10.1109/ICASSP39728.2021.9414177.

[2] X. Le, H. Chen, K. Chen, and J. Lu, “DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement,” in *Proceedings of Interspeech*, pp. 2811–2815 (Brno, Czech Republic) (2021 Aug.). https://doi.org/10.21437/Interspeech.2021-296.

[3] S. Braun and I. Tashev, “Data Augmentation and Loss Normalization for Deep Noise Suppression,” in A. Karpov and R. Potapova (Eds.), *SPECOM 2020: Speech and Computer*, Lecture Notes in Computer Science, vol. 12335, pp. 79–86 (Springer, Cham, Switzerland, 2020). https://doi.org/10.1007/978-3-030-60276-5_8.

[4] H. Dubey, V. Gopal, R. Cutler, et al., “ICASSP 2022 Deep Noise Suppression Challenge,” *arXiv preprint arXiv.2202.13288* (2022). https://doi.org/10.48550/arXiv.2202.13288.

[5] Apple, “*Capturing Stereo Audio From Built-In Microphones*,” https://developer.apple.com/documentation/avfaudio/avaudiosession/capturing_stereo_audio_from_built-in_microphones (2022).

[6] J. Clover, “Apple Now Has More Than 1.8 Billion Active Devices Worldwide,”. https://www.macrumors.com/2022/01/27/apple-1-8-billion-active-devices-worldwide/ (2022 Jan.).

[7] Microsoft, “Microphone Array Geometry Property,” https://learn.microsoft.com/en-us/windows-hardware/drivers/audio/microphone-array-geometry-property (2021 Dec.).

[8] Roland Corporation, “CS-10EM Binaural Microphones/Earphones,” https://www.roland.com/us/products/cs-10em/ (2022).

[9] Sennheiser Electronic, “AMBEO Smart Headset,” https://en-us.sennheiser.com/in-ear-headphones-3d-audio-ambeo-smart-headset (2022).

[10] A. S. Master, L. Lu, H.-M. Lehtonen, et al., “Dialog Enhancement via Spatio-Level Filtering and Classification,” presented at the *149th Convention of the Audio Engineering Society* (2020 Oct.), e-Brief 637. https://doi.org/10.17743/aesconv.2020.978-1-942220-33-6.

[11] I. Stewart, “What is Mid/Side Processing?,” *iZotope* (2022 Jul.). https://www.izotope.com/en/learn/what-is-midside-processing.html.

[12] D. Keller, “Mid/Side Mic Recording Basics,” https://www.uaudio.com/blog/mid-side-mic-recording/. (Date of Access: Nov. 10, 2022)

[13] A. Master, *Stereo Music Source Separation via Bayesian Modeling*, Ph.D. Dissertation, Stanford University, Stanford, CA (2006 Jun.).

[14] A. Øland and R. Dannenberg, “Loudness Concepts and Pan Laws,” http://www.cs.cmu.edu/∼music/icm-online/readings/panlaws/panlaws.pdf (Date of Access: Nov. 10, 2022).

[15] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847 (2004 Jun.). https://doi.org/10.1109/TSP.2004.828896.

[16] H.-S. Choi, S. Park, J. H. Lee, et al., “Real-Time Denoising and Dereverberation With Tiny Recurrent U-Net,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5789–5793 (Toronto, Canada) (2021 Jun.). https://doi.org/10.1109/ICASSP39728.2021.9414852.

[17] Y. Hu, Y. Liu, S. Lv, et al., “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,” in *Proceedings of the Interspeech*, pp. 2472–2476 (Shanghai, China) (2020 Oct.). https://doi.org/10.21437/Interspeech.2020-2537.

[18] Y. Liu and D. Wang, “Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2092–2102 (2019 Dec.). https://doi.org/10.1109/TASLP.2019.2941148.

[19] J. T. Geiger, P. Grosche and Y. L. Parodi, “Dialogue Enhancement of Stereo Sound,” in *Proceedings of the 23rd European Signal Processing Conference*, pp. 869–873 (Nice, France) (2015 Aug.). https//doi.org/10.1109/EUSIPCO.2015.7362507.

[20] 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 (2019 Jul.). https://doi.org/10.17743/jaes.2019.0032.

A.1 STEREO-POLAR DATA

In the CMSS processing system described above, it was noted that the SISSD estimates $\Theta_1(b, t)$ and $\Phi_1(b, t)$ are obtained from STFT data. It was also noted that the STFT tiles that contain only a particular target source lead to particular estimates. To develop a fuller understanding of these cases, the authors presently describe a mapping for the data in a typical stereo STFT representation to an alternative form termed *stereo-polar coordinates*. Recall that stereo STFT data is typically represented as the real and imaginary components or magnitude and phase for each channel. The estimation process described in the SISSD first transforms it into an alternative, *stereo-polar coordinates representation* (SPCR), which allows for parameter estimation as described in [10]. When including all four values below, SPCR is convertible to and from a conventional stereo STFT representation and includes values for the following for each STFT tile, where L and R are the STFT representations of the left and right channels.

$$
\begin{aligned}
U &= \sqrt{|L|^2 + |R|^2} \\
\theta &= \arctan\left(\frac{|R|}{|L|}\right) \\
\phi &= \angle\left(\frac{L}{R}\right) \\
\psi &= \angle L - \phi \sin^2\theta \\
&= \angle R + \phi \cos^2\theta.
\end{aligned}
$$

The first parameter, U, is the *combined channel magnitude* and may be thought of as the data for a combined mono spectrogram. The second parameter, θ, is the *mapped ILD*, which ranges from 0 (for L much greater in magnitude) to $\pi/2$ (for R much greater in magnitude). Unlike alternatives (e.g., Eq. (22) in [15]), it is strictly bounded. Together, these two quantities describe the *stereo-polar magnitude* of an STFT tile. The third parameter, ϕ, describes the IPD from $-\pi$ to π radians, and the fourth, ψ, the *base phase* also from $-\pi$ to π radians; together these describe *stereo-polar phase.* The choice of ψ here with respect to ϕ follows a similar convention as for Ψ_1 above with respect to Φ_1 in the generalized target signal model. Altogether, the four quantities above are termed the SPCR. L and R can be reconstructed from the SPCR by calculating:

$$L = U\cos\theta\exp(i(\psi + \phi\sin^2\theta))$$

$$R = U\sin\theta\exp(i(\psi - \phi\cos^2\theta)).$$

Similarities and differences are noted between the quantities $(\Theta_1, \Phi_1, \Psi_1, |S_1|)$ and (θ, ϕ, ψ, U). The values $|S_1|$ and Ψ_1 represent the magnitude and phase of a monaural target source (which can vary across each bin in STFT space), and Θ_1 and Φ_1 are its mixing parameters that characterize its presence in two channels (these parameters vary only if the source moves, is reverberant, or is mixed with inter-channel delay). The values (θ, ϕ, ψ, U), however, are detected for *each* STFT tile and may or may not coincide with the target source values depending on whether a given STFT tile is dominated by the target source, by interferers, or some combination. (See, e.g., [15].) If a given tile is dominated by the target source, it is easy to see by substitution of the L and R values for the generalized source model in Sec. 2 that (θ, ϕ, ψ, U) will equal $(\Theta_1, \Phi_1, \Psi_1, |S_1|)$; this fact forms the basis for the SISSD estimation described in [10], which analyzes distributions on the values (θ, ϕ, U) to calculate $\Theta_1(b, t)$ and $\Phi_1(b, t)$.

THE AUTHORS

Aaron Master

Lie Lu

Nathan Swedlow

Aaron Master received a B.S.E.E. from the University of Rochester (NY) in 1999, B.Mus. from the Eastman School of Music in 1999, M.Phil. in Engineering from the University of Cambridge (UK) in 2000, and Ph.D. in Electrical Engineering from Stanford University in 2006. From 2006 to 2013, he worked as a Research Engineer and UX Director at SoundHound Inc., where he was a lead inventor of technologies allowing combined query-by-humming and automatic content recognition (ACR), instant-response ACR, automatically synchronized lyrics, and song popularity prediction. Apps he managed received awards from the *New York Times*, *Time* magazine, CNET, and Billboard. From 2013 to 2022, Dr. Master served as Manager and Senior Manager of Sound Technology at Dolby Laboratories, where he led work on source separation for dialog enhancement. Additional research interests include spatial audio, music information retrieval, and human perception. Dr. Master is first author of over 30 peer-reviewed papers and patents.

•

Lie Lu received his B.S. and M.S. degrees in Electrical Engineering from Shanghai Jiao Tong University, China, in 1997 and 2000, respectively, and a Ph.D. degree from Delft University of Technology, The Netherlands, in 2009. He is currently a Senior Member of Research Staff at Dolby Laboratories in San Francisco. Before that, he was with Microsoft Research Asia, Beijing, China, from 2000 to 2010, first in the Media Computing group and then in the Speech group. His current research interests include machine learning, signal processing, content-based audio analysis, and music information retrieval. He has published over 60 papers in scientific journals and leading conferences in the area of audio and speech processing and is an author on over 70 issued or pending patents.

•

Nathan Swedlow is a neuroscientist and musician who graduated from Oberlin College and Conservatory in 2015. Through perceptual and physiological research, he works to uncover how technology impacts multi-modal sensory experiences. Nathan has worked at Dolby Laboratories since 2016. Outside of his work at Dolby, Nathan is an active musician and audio engineer.