# Evaluating source separation algorithms with reverberant speech

Michael I. Mandel, *Member, IEEE,* Scott Bressler, Barbara Shinn-Cunningham,
Daniel P. W. Ellis, *Senior Member, IEEE*

***Abstract*—This paper examines the performance of several source separation systems on a speech separation task for which human intelligibility has previously been measured. For anechoic mixtures, automatic speech recognition (ASR) performance on the separated signals is quite similar to human performance. In reverberation, however, while signal separation has some benefit for ASR, the results are still far below those of human listeners facing the same task. Performing this same experiment with a number of oracle masks created with *a priori* knowledge of the separated sources motivates a new objective measure of separation performance, the DERTM (Direct-path, Early echo, and Reverberation, of the Target and Masker), which is closely related to the ASR results. This measure indicates that while the non-oracle algorithms successfully reject the direct-path signal from the masking source, they reject less of its reverberation, explaining the disappointing ASR performance.**

## I. Introduction

It is important to measure how well source separation algorithms perform in order to compare and improve them. Typically, however, this performance is measured in terms of the similarity between the original source waveforms and the waveforms estimated from a mixture. When separating sources from an instantaneous overdetermined mixture (i.e. a mixture with more microphones than sources and no reverberation or delays), perfect reconstruction of each source should be attainable. When separating sources from reverberant and underdetermined mixtures, however, perfect reconstruction is impossible because of the information lost in the mixing process. Thus, measuring the fidelity of the waveform estimation makes unreasonable expectations of a separation algorithm.

It is more informative to compare the original and estimated signals in a representation that is more relevant to the intended application of the source separator than the raw waveform [1]. For example, if a separator is to be used as the front-end for an automatic speech recognition system, its performance should be measured in terms of the increase in performance of the end-to-end system when it is in place, not in terms of an intermediate quantity that only indirectly affects the system's performance such as waveform reconstruction error. Similarly, if a separator is to be used in a hearing aid application, its performance should be measured in terms of the increase in *intelligibility* it provides to listeners.

Because human speech recognition accuracy is still well beyond that of automatic speech recognition (ASR) in many conditions, human speech recognition performance can be thought of as an upper bound on the performance of automatic recognizers. While human intelligibility has been used to evaluate the quality of ground truth masking-based source separation [2], [3], such evaluations are expensive and time consuming and must be re-run for every variant of a source separation algorithm. ASR performance, on the other hand, requires only computational resources. It has been used to evaluate some of the earliest underdetermined source separation systems (e.g., [4]).

ASR and intelligibility both provide a coarse measure of the performance of source separation systems, however, because they only provide estimates at the word or phoneme level. When developing source separators it is useful to have a metric that operates at a finer resolution in both time and frequency, thus allowing greater insight into how and why an approach works or fails. This has motivated the use of other *objective* measures of source separation performance. While such metrics, e.g., the perceptual evaluation of speech quality (PESQ) and various signal-to-distortion metrics, can be informative, they have not been shown decisively to correspond well with intelligibility of signals for human subjects.

PESQ is, as its name says, a metric designed to predict subjective judgments of speech *quality*. It was originally designed for use in telecommunications, but has recently been shown to correspond well will subjective judgments of speech quality for speech enhancement and separation systems [5]. It has also been used to predict ASR performance [6]. Quality is a property of speech that corresponds to its realism and naturalness, characteristics that are not necessary for intelligibility. [3] and [7] have shown that a frequency-dependent gating or modulation of noise, which has low quality, can be highly intelligible. Thus, while ASR performance can be predicted by quality to some extent, the relationship is imperfect and indirect [8], [9].

Other metrics that are popular for evaluating speech separation performance are based on some form of signal-to-noise ratio (SNR). While they are more straightforward to calculate, it is not clear whether they are meant to predict intelligibility or quality. As measures of intelligibility, these metrics are flawed, since signal energy does not correlate consistently with intelligibility [1]. For example, low speech frequencies tend to be relatively intense, independent of spectral and temporal

M. Mandel is with the Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, QC, H3C 3J7 Canada. S. Bressler and B. Shinn-Cunningham are with the Department of Cognitive and Neural Systems, Boston University, Boston, MA, 02215 USA. D. Ellis is with the Department of Electrical Engineering, Columbia University, New York, NY, 10027 USA e-mail: mandelm@iro.umontreal.ca, scbressler@verizon.net, shinn@cns.bu.edu, dpwe@ee.columbia.edu

Manuscript received November 16, 2009; revised February 16, 2010.

content, but do not affect intelligibility as much as mid and high frequencies. Similarly, in normal speech higher frequencies tend to have low intensity, but their deletion can diminish intelligibility significantly. Another problem with many such measures is that they do not provide a single figure of merit, but instead separately account for target energy that is lost and noise that is not eliminated (e.g., [10], [11]) complicating comparisons of different algorithms and making it difficult to determine how a given algorithm is failing.

This paper explores the performance of a number of state-of-the-art source separation systems in terms of the improvement they give to automatic speech recognizers and to human listeners. Sec. II-C proposes a number of models of the performance of these separators in terms of novel oracle-based source separators that preserve or reject different parts of the target and masker signals. Secs. III and IV show that while the non-oracle systems improve ASR performance substantially for anechoic mixtures, they only improve it slightly for reverberant mixtures. One of the oracle models very closely follows the performance of the source separators in these experiments and in subsequent investigations with other evaluation metrics in Secs. V and VI.

Sec. V examines these separations through the lens of the novel DERTM metric, which is related to this family of oracle-based models and measures the change in energy of the Direct-path, Early echos, and Reverberation of the Target and Masker signals produced by a mask-based source separator. DERTM distinguishes between different types of separation errors and further supports the fit of this model, implying that the source separators successfully reject the direct-path signal from interfering sources, but fail to reject reverberation from them. DERTM also detects that one separation algorithm, MESSL, breaks down at very low target-to-masker ratios because of a failure of its localization initialization. This breakdown is not detected by the Perceptual Evaluation of Speech Quality (PESQ) [12, Sec. 10.5.3.3] or the $\text{SNR}_{\text{Li}}$ metric (our designation) [13], which are described along with other objective separation metrics in Sec. VI.

## II. Shared experimental details

The experiment in this paper is based on the human spatial hearing experiment of [14], but only includes a subset of the conditions tested in that paper. Sec. II-A describes the original experiment, and Sec. II-B describes the experiment that we carried out based on it. Note that while all of the algorithms compared in this paper are able to separate underdetermined mixtures, because the original psychoacoustic experiment used a determined mixture with two sources and two microphones, we follow the same procedure here.

### A. Details of original experiment

The experiment of [14] examined the intelligibility of a target utterance in the presence of a masking utterance in reverberant and anechoic conditions while varying the cues that a listener could use to separate the two streams. All utterances were spoken by the same person, so no speaker-dependent cues were available to the listeners. The cues that were available included the linguistic content of the utterances, spatial location, pitch, and timing.

The target utterance was a string of five digits, “one” through “nine”, “oh”, and “zero”. The masking utterance was a sentence from the TIMIT corpus, i.e., a grammatically correct sentence. The target speaker was always located in front of the listener, while the masking speaker was either located in front of the listener or at $90^\circ$ to their right. In certain conditions, the speaker’s voice was modified to be a single pitch throughout the utterance. The digit-string target always started after the masker sentence, so that differences in source onset could be used to identify the target. Subjects were tested in all possible combinations of intonation and spatial conditions.

Unlike many recent studies of speech-on-speech masking, this task greatly limited the subjects’ uncertainty about which source was the target and which was the masker (e.g., in contrast to [15]–[19]), essentially solving the problem of source selection (see [20]). Instead, listeners should have been able to understand the target if they could separate it from the masker. This made these results more directly relevant to our own separation algorithm evaluations than most other studies.

The results of this experiment indicated that the intelligibility of speech was qualitatively different in anechoic and reverberant conditions. In all conditions masking was maximal when intonation and spatial cues were the same in the two utterances. In reverberation, where segregation cues are weakened, differences in either of these cues provided a release from masking and differences in both of these cues combined into an even larger release from masking. This was not the case in the anechoic condition where segregation was much easier. In this case, the effect of spatial separation could be explained by energetic release, i.e., the signal at the ear with the higher SNR predicted all anechoic binaural conditions perfectly. Thus in anechoic conditions, pitch helped humans segregate sources, but location did not help beyond the energetic differences it induced in different configurations.

Generally, in reverberant conditions, the curve of intelligibility vs target-to-masker ratio (TMR) was shifted towards higher TMRs and had a more shallow slope than in anechoic conditions (see Figs. 1 and 2 for these curves). This indicates that the target was less intelligible in a reverberant mixture, but that an incremental change in relative masker level had a smaller effect on intelligibility in reverberation (i.e., that the TMR was less directly responsible for performance limitations than in anechoic conditions).

### B. Details of current experiment

The experiments performed in this paper used the same utterances and the same impulse responses as [14]. The impulse responses were recorded in a classroom of size $5 \times 9 \times 3.5$ m (the same classroom as [21]) with the source 1.2 m from a KEMAR dummy head. They have a reverberation time of approximately 550 ms and have a direct-to-reverberant ratio of approximately 10 dB, computed as the ratio of the energy in the first 9.6 ms of the impulse response to that in the rest of the impulse response. One binaural impulse response was recorded at each of a number of azimuths in the horizontal

TABLE I
THE SIX ORACLE MASKS USED IN THESE EXPERIMENTS. + INDICATES THAT THE MASK TREATS A SIGNAL COMPONENT AS DESIRABLE, − INDICATES THAT THE MASK TREATS IT AS UNDESIRABLE. SIGNAL COMPONENTS ARE: TARGET DIRECT-PATH, TARGET REVERBERATION, MASKER REVERBERATION, AND MASKER DIRECT-PATH. NOTE THAT EARLY ECHOES ARE INCLUDED IN REVERBERATION.

| Name | Type | Tgt DP | Tgt Rev | Msk Rev | Msk DP |
|---|---|---|---|---|---|
| DP-Oracle | Binary | + | − | − | − |
| Oracle | Binary | + | + | − | − |
| OracleAllRev | Binary | + | + | + | − |
| DP-Wiener | Wiener | + | − | − | − |
| Wiener | Wiener | + | + | − | − |
| WienerAllRev | Wiener | + | + | + | − |

plane; however, following [14], we only use the two recorded at $0^\circ$ and $90^\circ$. The impulse responses were recorded at a 50 kHz sampling rate, and then downsampled to 25 kHz, the sampling rate of the utterances.

There were some differences between our experiment and that of [14], however. Because localization is central to separation in all of the separation systems under evaluation, we did not evaluate them on the collocated conditions. Unlike the human subjects, the algorithms had no prior knowledge of the locations of the sources before the separation. We also only evaluated mixtures using naturally intoned speech; we did not use the monotonized speech.

The 200 unique target utterances were split into three sets, a training set of 75 utterances, a tuning set of 50 utterances, and a test set of 75 utterances. All three sets used the same two pairs of binaural room impulse responses and could use the same masking utterances. The training set consisted of 64 mixtures at each of 8 TMRs spaced evenly between $-1.7$ dB and 45 dB. The tuning and test sets consisted of 64 mixtures at each of 10 TMRs spaced evenly between $-35$ and 25 dB. The training mixtures were randomized at each TMR, but the 64 random mixtures in the tuning set were the same at each TMR, as were the 64 random mixtures in the test set. Thus, there were a total of 512 utterances in the training set and 640 utterances in each of the tuning and test sets.

As in the human experiment, the target utterances were always shorter than the maskers, and silence was inserted before and after them to center them in their respective maskers. All utterances were normalized to the same RMS level before being spatialized with the binaural room impulse responses and mixed at the appropriate TMR. All mixtures had dither added to them to create a constant noise floor in the inserted silence. The dither just covered the recording noise floor in the louder ear at the highest TMR of an utterance that had been passed through the anechoic impulse responses.

### C. Oracle separations

The oracle binary mask, also known as the ideal binary mask, has been proposed as an upper bound on the performance of source separation algorithms that generate binary time-frequency masks [13]. It is "ideal" both in being optimal in terms of signal-to-interference ratio and in being unattainable from real-world mixtures. This is because it is created using knowledge of the original component signals before they are combined to make a mixture. In this paper, we compare six different oracle separation algorithms. All of these oracle masks are constructed based on the ratio of desirable energy to undesirable energy at each time-frequency point. They differ in their classification of the desirability of direct-path and reverberant signals from the target and masker sources, and in whether they create binary or continuous (Wiener filter) masks. See Tab. I for an indication of the classification of these signals in each mask.

The final reverberated signal can be decomposed into its direct-path, early echo, and reverberant portions by convolving different parts of the impulse response with the original anechoic signal, or, equivalently, by settings various parts of the impulse response to 0. The direct path portion of each impulse response was generated from the full reverberant impulse response by setting to 0 all samples more than 9.6 ms from its beginning. Similarly, the early echo portion set to 0 all samples less than 9.6 ms and more than 32 ms from the beginning, and the reverberant portion set to 0 all samples less than 32 ms from the beginning. Note that for the purposes of constructing the oracle masks, early echoes were considered to be part of the reverberation, as they did not make an appreciable difference on performance.

The DP-Oracle mask is a binary mask that only considers the target direct-path signal to be desirable. The Oracle mask is the same, but also considers the reverberation of the target signal to be desirable. The OracleAllRev mask additionally considers the reverberation of the interfering signal to be desirable. This mask is proposed as an idealized comparison for the algorithmic separations described later in this paper, which tend to successfully reject the direct-path signal from the interfering source, but have more difficulty rejecting its reverberation. For each binary mask, there is a corresponding Wiener filter mask computed as the ratio of the desirable energy to the total energy at each time-frequency point.

### D. Algorithmic separations

These experiments involve our own separation system, Model-based EM Source Separation and Localization (MESSL) [22] and three other well known source separation algorithms: DUET [23], the algorithm of [24], which we refer to as Sawada, and the algorithm of [25], which we refer to as Mouba. We implemented all of these systems ourselves as we did not have access to the authors' original code.

Model-Based EM Source Separation and Localization (MESSL) [22] jointly separates and localizes sound sources by probabilistically modeling the interaural parameters of each source. The interaural level difference (ILD) is modeled as a diagonal-covariance Gaussian with frequency-dependent mean and standard deviation. The interaural time difference (ITD) is modeled as a hidden multinomial that is only observed in the interaural phase difference (IPD), which is modeled with a mixture of diagonal covariance Gaussians, each with frequency-dependent means and standard deviations. The hidden ITD variables constrain the IPD observations to be approximately linear with frequency. MESSL treats the interaural level and

phase differences as independent variables, so they only interact through the mask that each generates. Additionally, all points in the spectrogram are assumed to be independent of one another when conditioned on the interaural parameters (ILD, ITD, and IPD). Typically, only the IPD parameters are initialized; this initialization is set by a cross-correlation estimate of the source ITD. Other parameters (e.g., ILD parameters) are first estimated from the regions that the initial IPD selects as relevant. Two versions of MESSL were evaluated here. The first, denoted MESSL-G, used an explicit "garbage source" model to absorb reverberation and a prior on the ILD parameters to more reliably estimate them in reverberation. The second, denoted MESSL-WW, did not use the garbage source or ILD prior.

Two-stage frequency-domain blind source separation [24], which we refer to as Sawada, is a combination of ideas from model-based separation and independent component analysis (ICA) that can separate underdetermined mixtures. In the first stage, blind source separation is performed on each frequency band of a spectrogram separately using a probabilistic model of mixing coefficients. In the second stage, the sources in different bands are unpermuted using k-means clustering on the posterior probabilities of each source and then refined by matching sources in each band to those in nearby and harmonically related bands. The first stage encounters problems when a source is not present in every frequency and the second encounters problems if sources' activities are not similar enough across frequency.

Like MESSL, the algorithm of [25], which we refer to as Mouba, uses EM clustering to separate sources from binaural recordings. This algorithm needs access to certain coefficients describing the relationship between ILD, ITD, and azimuth, which can be extracted offline from head-related transfer functions. It is not particularly sensitive to the exact values of these coefficients, however, so the same values generally work for different heads. Using these coefficients, the algorithm maps the ILD at each point in the spectrogram to an azimuth, with which it disambiguates each IPD-to-ITD mapping. The ITD is then mapped to azimuth at each spectrogram point and these azimuth values (after weighting by the energy at that spectrogram point) are clustered using a Gaussian mixture model. The means of the Gaussians are the estimated source locations and the posterior probabilities of each azimuth coming from each Gaussian are used to construct a spectral mask.

The Degenerate Unmixing Estimation Technique (DUET) [11], [23] creates a two-dimensional histogram of the interaural level and time differences observed over an entire spectrogram. It then smooths the histogram and finds the $I$ largest peaks, which should correspond to the $I$ sources. DUET assumes that the interaural level and time differences are constant at all frequencies and that there is no spatial aliasing, conditions that can be met to a large degree with free-standing microphones close to one another. With dummy head recordings, however, the ILD varies with frequency and the microphones are spaced far enough apart that there is spatial aliasing above approximately 1 kHz. Frequency-varying ILD scatters observations of the same source throughout the histogram as does spatial aliasing, making sources harder to localize and isolate.

While all four of these systems are compared in Sec. III, only Sawada and MESSL, the best performing algorithms, are compared after that to avoid clutter. No modifications were made to the algorithms; they were all required to localize the sources before separating them, if necessary.

## III. Automatic speech recognition

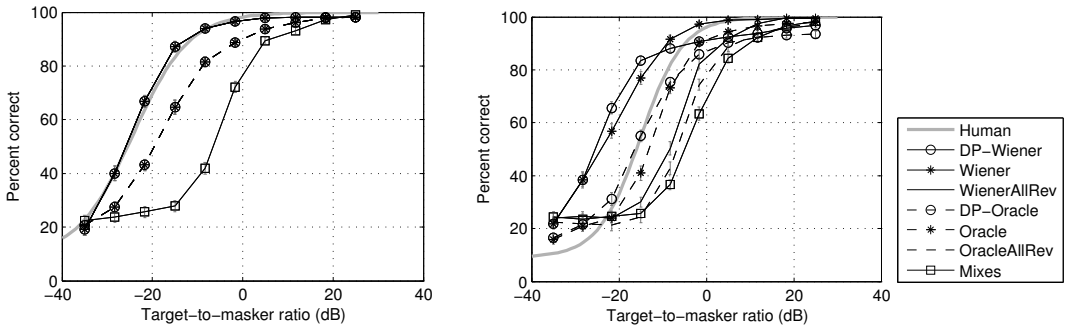
As described in e.g., [26]–[31], automatic speech recognition (ASR) systems have been incorporated into and used to evaluate source separation systems. Perhaps the earliest example of evaluating source separation with ASR was the work of [4]. While a speech recognition system can give a good idea of how well a source separator is doing, depending on the application, the ultimate goal might be presenting the separations to humans, in which case intelligibility would be paramount. If used carefully, however, ASR results can be a reasonable proxy for actual intelligibility measurements, which are costly and time-consuming to perform.

### A. ASR setup

While many papers describe a single recognizer with which a number of different separators are tested [32], [33], we instead train a different recognizer for each separation system. Speech recognizers can be easily "distracted" by artifacts of source separators that are not disruptive to intelligibility. This can be mitigated, however, by training a new recognizer on the output of each separator, because the recognizer is able to account for the particular artifacts of each separator in its statistical modeling. While this might not be feasible for a real-world system that selects different separation front-ends for different conditions, it should be feasible for a system with a dedicated source separator on the front end and it gives the "fairest" comparison of ASR performance in ideal situations. Moreover, we have found in pilot studies that recognizers trained with signals from multiple separators are generally good at recognizing speech from both, possibly at the cost of requiring a larger model.

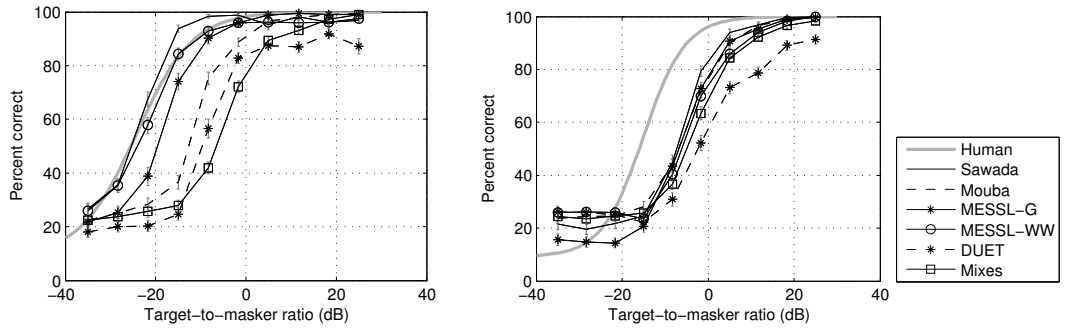
All of these experiments were conducted using the HTK speech recognition framework, and all of the recognizers had the same structure and number of parameters. Each separator reconstructed the waveform as best it could, and this waveform was fed into the recognition system; as a result, missing data [27] was not accounted for. The recognizers modeled each of the 11 words separately with a linear 16-state HMM. The only grammatical constraint was that there must be five digits in each sentence. Each state modeled its output using a mixture of 8 Gaussians. The features used in the recognition process were 13-dimensional MFCCs including $C_0$, MFCC deltas, and MFCC double-deltas. Two different recognizers were trained for each algorithm, one on only anechoic separations, and one on only reverberant separations. A recognizer trained on both anechoic and reverberant separations performed comparably.

As described in Sec. II-B, the training, test, and tuning sets used separate sets of target utterances to prevent memorization of the targets. They did use the same maskers, but because these were never recognized, they should not affect the results. The mixtures were randomized in the training set at different TMRs to train the recognizer with as many different contexts



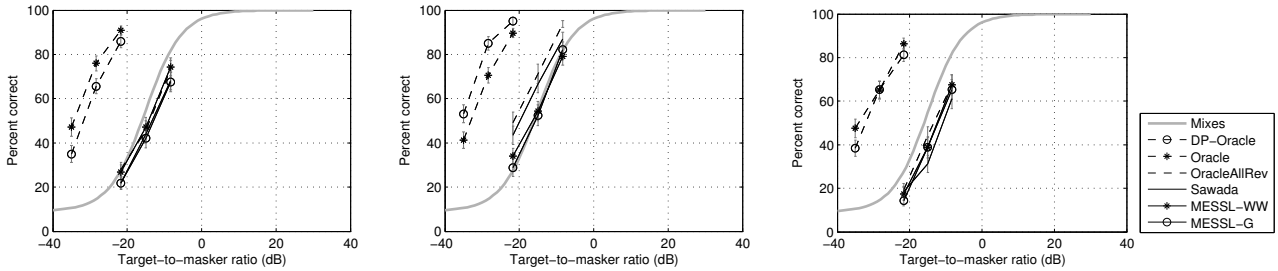

(a) Anechoic (b) Reverberant

Fig. 1. Automatic speech recognition performance as a function of target-to-masker after applying ground truth separation masks. Note that in (a) many of the lines lie exactly on top of each other. Error bars show 1 standard error.




(a) Anechoic (b) Reverberant

Fig. 2. Automatic speech recognition performance as a function of target-to-masker ratio after applying algorithmic separation masks.




(a) Subject 1 (b) Subject 2 (c) Subject 3

Fig. 3. Intelligibility of source separations as a function of target-to-masker ratio in reverberation for three subjects. The thick gray lines shows the intelligibility of unprocessed mixtures. Error bars show 1 standard error.

of each digit as possible. Even though there were only 75 original training utterances, the use of these different contexts should justify training on 512 separations.

While humans asked to transcribe 5 digits tend to make mostly substitution errors, this ASR system makes a combination of insertion and substitution errors. We thus had to evaluate it differently than humans are evaluated. Specifically, we used the fraction of digits that were correctly identified (COR) after the system's transcription and the true transcription were aligned using the minimum edit distance with equal weights.

### *B. Results*

The results of the ASR experiment for ground truth separations can be seen in Fig. 1. Human performance from [14] on the unprocessed mixtures is shown as the thick gray line and the performance of the recognizer trained and tested directly on the unprocessed mixtures is shown in the line with square markers. In Fig. 1(a), because there is no reverberation, the masks that treat reverberation differently are identical, and thus all of the binary masks lead to exactly the same performance and all of the Wiener masks lead to exactly the same performance. Note that the Wiener masks have an advantage of a few dB over the binary masks in terms of the 50% correct point on their curves, known as the speech reception threshold (SRT). The recognizers perform at close to human levels, particularly those trained with Wiener masks, and significantly better than recognizers trained and tested on the mixtures.

In reverberation, the results are slightly different. In Fig. 1(b), the difference between the Wiener and binary masks is still maintained. The Wiener masks significantly out-perform the

human subjects, achieving approximately the same performance that they do in anechoic conditions, while the binary masks perform comparably to the humans, although also similarly to their performance in anechoic conditions. There is only a slight difference in performance between the masking systems that include and exclude the target reverberation. But, notably, including the masker's reverberation significantly decreases performance for both mask types. The WienerAllRev mask performs slightly better than the mixtures, but the OracleAllRev mask only provides as much information to the recognizer as the unseparated mixtures.

The results of the ASR experiment for algorithmic separations can be seen in Fig. 2. Once again, human performance is indicated by the thicker gray line and the performance on unprocessed mixtures is indicated by the line with square markers. Fig. 2(a) shows that the performance in anechoic conditions of Sawada, MESSL-G, and MESSL-WW is quite close to human performance. Mouba is able to achieve some improvement, but DUET only achieves improvement at TMRs close to 0 dB.

In reverberation, however, ASR performance on the separations is quite close to that of the mixtures, i.e. the separation systems only help a little bit. Sawada, Mouba, and both MESSL systems are clustered together, with a decrease in SRT of perhaps 3–4 dB over the mixtures. DUET actually performs worse than the mixtures, meaning that it makes the target speech less recognizable than before it was applied.

We examine some possible causes of this lack of improvement in subsequent sections for these same separations. Sec. V examines the effects of some of these masking algorithms on the energy in the direct-path, early echo, and reverberation components of the target and masker sources. Sec. VI examines objective measures of speech intelligibility and quality.

## IV. Intelligibility of Separations

In addition to running automatic speech recognition on the resulting separations, we also directly measured the intelligibility of the reverberant separations for human listeners. While this experiment was more limited than the experiment in [14] in both the number of subjects and number of TMRs tested, it provides another interesting comparison of the separation algorithms. Because of the similarity of the binary and Wiener ground truth systems, we will only examine the performance of the three binary ground truth separators. Because of their better performance, we will only compare Sawada, MESSL-G, and MESSL-WW in this and subsequent sections. And because it is a more difficult and realistic task, we will also only examine the performance of these systems on the reverberant mixtures.

Three native English-speaking, college-age, female subjects were recruited to perform the follow-up experiment. All subjects had their audiograms measured to confirm their hearing thresholds were within 15 dB of normal thresholds for all frequencies between 250 and 8000 Hz. All subjects gave written informed consent to participate in the experiments, as approved by the Charles River Institutional Review Board.

Each subject was presented with separations from six algorithms (the three binary masking algorithms, Sawada, MESSL-G, and MESSL-WW) at three different TMRs. Each subject was presented with the same 64 mixtures for each algorithm at each TMR. The specific values of the TMRs used for each algorithm were selected based on preliminary piloting experiments. The signals presented to the subjects were each algorithm's best estimate of the target digit string in a reverberant mixture. These signals were presented to the subjects over ER-1 insert-ear headphones (Etymotic Research, Inc.) in a sound isolated booth. Subjects were asked to transcribe the five digits they heard and were required to enter five digits on every trial with attention to order, and were instructed to guess if they were unsure. Feedback was not provided.

Due to the large number of audio presentations (1152 in total), the experiment was divided into two sessions occurring on two different days. Each session tested a subset of three of the six source separation algorithms. To mitigate any potential learning effects, two of the subjects tested the Sawada, DP-Oracle, and OracleAllRev algorithms in the first session, and the Oracle, MESSL-G, and MESSL-WW algorithms in the second session. The other subject was presented with these two groups of algorithms in reverse order. Additionally, the order of the combination of algorithm and TMR within session was randomized from trial to trial. The proportion of correctly reported digits was computed from these results.

### A. Results

The results of these intelligibility experiments can be seen in Fig. 3 for each of the three subjects and each algorithm. As can be seen from the individual plots, the subjects are generally in agreement with each other and two results are clear. The first is that the DP-Oracle and Oracle masks create separations that are substantially more intelligible than the other four masks, but that are quite comparable to one another. The second is that the algorithmic separations have intelligibilities that are quite similar to the OracleAllRev mask, and quite similar to the intelligibility of the unprocessed mixtures. These results are qualitatively similar to the automatic speech recognition results, although quantitatively different.

## V. DERTM: Attenuations of Six Signal Components

We now describe a novel evaluation metric for mask-based separations, which we refer to as DERTM: Direct-path, Early echoes, and Reverberation of Target and Masker. As mentioned in Sec. II-C, a reverberant mixture of two sources contains at least six signals of interest. Each of the two sources contains a direct-path, an early echo, and a reverberant portion. The DERTM analysis measures the attenuation of each of these six signals by a given mask, revealing differences in a mask's effect on each. This is possible because all of the masks under evaluation contain values between 0 and 1, meaning that they only delete energy, they do not add it[1]. Furthermore, because the mixtures are synthesized, we have access to the original

[1]Note that this does not account for the potential "transformation" of masker energy into target energy, as in [3], although such a transformation is probably less likely to happen in pure speech-speech mixtures than in the speech-noise mixtures those authors examined

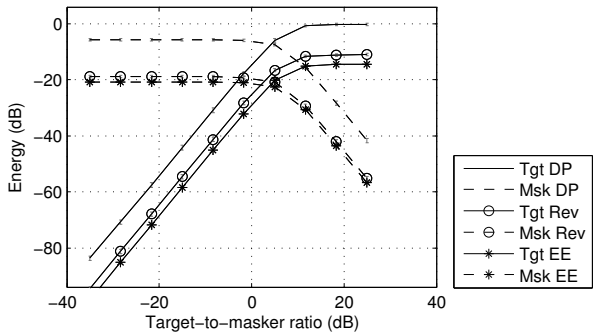



Fig. 4. Initial levels of the six different types of energy: target direct-path, target reverberation, masker direct-path, masker reverberation. Error bars show 1 standard error.

versions of these six signals for every mixture. It is informative to compare this analysis for each algorithm to the ASR and intelligibility results in Secs. III and IV.

The results of this analysis are shown in Fig. 5 for each algorithm as the attenuation of each of these signals relative to the initial levels shown in Fig. 4. Examining the oracle masks in the top row, it can be seen that each preserves most the energy it considers desirable while not entirely canceling undesirable energy. Oracle masks are not able to perfectly pass or attenuate signals because they can only apply a single weighting to each time-frequency point. All of the energy in a given point, whether desirable or not, must be treated identically.

The DP-Oracle mask is able to attenuate the other signals by a wide margin relative to the target signal's direct-path component, especially those of the masker. The early echo energy from the target is attenuated by only 5 dB because it shares many of the same time-frequency points as the direct-path energy. The late reverberation from the target is attenuated by 10 dB and the fact that it is not attenuated more indicates the presence of overlap-masking [34]. The three signal components from the masking source are attenuated significantly, especially when they have larger initial energies. The direct-path is attenuated most, possibly because it is initially louder than the other signals.

The Oracle mask classifies the reverberation from the target source as desirable and it attenuates this reverberation much less than the DP-Oracle mask, although it does not preserve it as well as it preserves the direct-path. It also attenuates the masking source less than the DP-Oracle mask does, especially at higher TMRs. The OracleAllRev mask performs quite differently from the two other oracle masks. It does a much better job of preserving all five of its target signals, but also does a much worse job of attenuating the direct-path masker signal. Instead of a consistent fall off with decreasing TMR, its performance is relatively constant at low TMRs.

In general shape, Sawada et al.'s algorithm performs very similarly to OracleAllRev under this metric. Sawada et al.'s algorithm attenuates all of the sources slightly more than OracleAllRev, although it does a better job attenuating target and masker reverberation and early echoes relative to target direct-path. It consistently rejects the direct-path portion of the masker signal. MESSL-WW performs similarly for high TMRs, but clearly begins to break down for TMRs below $-10$ dB. The convergence of the six lines for very low TMRs is perhaps the clearest indication of such a breakdown over all the metrics that will be discussed in this paper. It indicates that MESSL is producing a mask that attenuates the entire mixture by 6 dB, i.e. is not selective at all. For higher TMRs, its performance is still not quite as good as Sawada's; in particular, it does not reject as much reverberation from either source or the direct-path of the masking source as Sawada's algorithm.

MESSL-G also fails for very low TMRs, but it does so in a different manner than MESSL-WW. Instead of attenuating all six components by 6 dB or so, it actually attenuates all of the signals by up to 20 dB except for the masker direct-path. Examining the problem in closer detail, what appears to happen is that a failed localization results in both of the non-garbage sources modeling the masker source and the garbage source absorbs the target.

## VI. Objective measures

This section examines the performance of these systems under a number of standard objective measures of speech quality and intelligibility, in addition to a novel measure proposed here. We performed these measurements both with and without a pre-emphasis high pass filter, but no significant differences were noted, so this section only reports results that do not use pre-emphasis.

### A. BSS_EVAL metrics

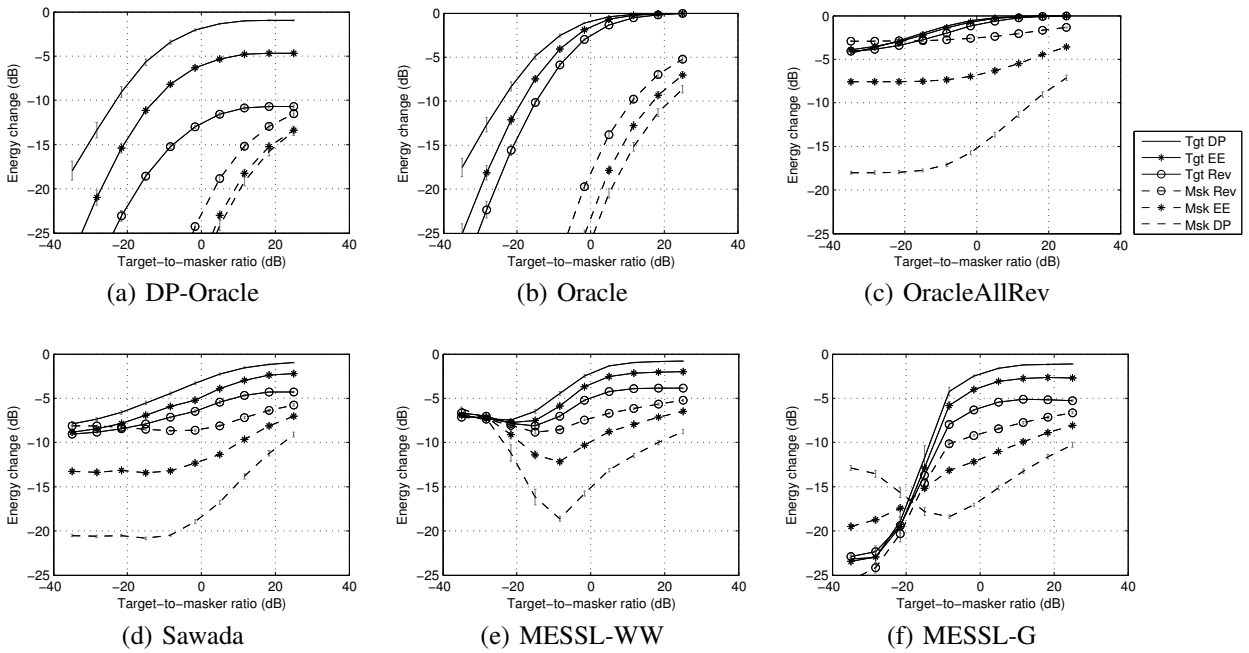
The first objective measures of separation performance come from the Signal Separation Evaluation Campaign [35] and the BSS_EVAL toolbox [36], which measure certain energy ratios of interest while properly accounting for scaling and convolution with short impulse responses. Specifically, there are three ratios of interest, the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR). These three ratios are defined in relation to three time-domain waveforms created by projecting the estimated signal onto the space of the original signals: the target signal embedded in the estimate, $s_{\text{target}}(t)$, the error due to interference, $e_{\text{interf}}(t)$, and the error due to artifacts, $e_{\text{artif}}(t)$.

To describe these signals mathematically, some notation is required. Denote the set of original, anechoic signals as $\{s_i(t)\}$, the anechoic target signal as $s_j(t)$, and the estimated target signal as $\hat{s}_j(t)$. Define the projection operator $P(x, \{y_i\}, \tau_{\text{max}})$ to be the projection of signal $x$ onto versions of the signals $\{y_i\}$ shifted by every integral number of samples up to $\tau_{\text{max}}$. Thus for signals $T$ samples long, this projection uses a matrix that is $T \times T$ and rank $I\tau_{\text{max}}$ where $I$ is the number of signals $\{y_i\}$. This projection results in another length-$T$ signal. The three signals of interest are then defined as

$$s_{\text{target}}(t) = P(\hat{s}_j, s_j, \tau_{\text{max}}) \quad (1)$$

$$e_{\text{interf}}(t) = P(\hat{s}_j, \{s_i\}, \tau_{\text{max}}) - P(\hat{s}_j, s_j, \tau_{\text{max}}) \quad (2)$$

$$e_{\text{artif}}(t) = \hat{s}_j - P(\hat{s}_j, \{s_i\}, \tau_{\text{max}}) \quad (3)$$




(a) DP-Oracle (b) Oracle (c) OracleAllRev

(d) Sawada (e) MESSL-WW (f) MESSL-G

Fig. 5. DERTM analysis of masking separations. Energy change due to mask attenuation of six different types of energy: target direct-path, early echoes, and reverberation; masker direct-path, early echoes, and reverberation. Shown for ground truth masks (top row) and algorithmic masks (bottom row). Error bars show 1 standard error.

and the three ratios of interest are defined as

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2} \tag{4}$$

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{5}$$

$$\text{SAR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2} \tag{6}$$

where $\|\cdot\|^2$ indicates the squared vector 2-norm, i.e. the sum of squares of all entries. To give some intuition about these definitions, $s_{\text{target}}$ is whatever part of $\hat{s}_j$ can be explained as a finite impulse response-filtered version of the target source, where the filter has $\tau_{\max}$ taps. The interference error is the part of $\hat{s}_j$ that can be explained by filtered versions of the other sources using filters of the same length. The artifact error is anything that cannot be attributed to these projections. In these experiments, $\tau_{\max}$ is 512 samples, which is approximately 21 ms at the sampling rate of 25 kHz. Thus, only the early echoes should be explained by the delayed versions of the sources, and later reverberation from both sources is counted as artifactual. Any other noise created by nonlinear processing, such as musical noise or burbling, is also counted as an artifact.

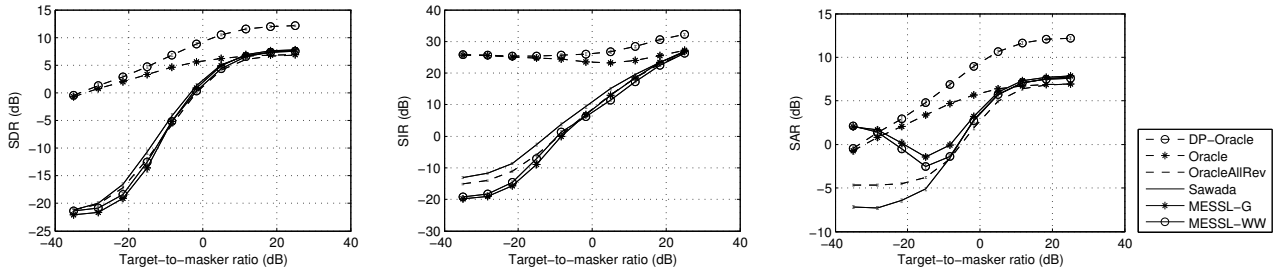
Fig. 6 shows the results of the separation systems using these metrics. The general TMR trend is evident in the plots of SDR and SIR except for the DP-Oracle and Oracle ground truth masks. Overall, the OracleAllRev mask performs quite similarly to the algorithmic separations, especially Sawada's. As evidenced from the SAR plot, MESSL's separations qualitatively change for TMRs below $-10$ dB, in agreement with the DERTM analysis in Sec. V. For TMRs above that point, the SDR performance of MESSL and Sawada are very close and significantly higher than OracleAllRev. Below that point, however, Sawada's algorithm maintains its performance while MESSL begins to fail. From the DP-Oracle and Oracle masks, it is clear that better performance is possible, although it is not clear that such performance is attainable without *a priori* knowledge of the signals, especially at very low TMRs.

Under the SIR metric, as seen in Fig. 6(b), Sawada's algorithm performs significantly better than MESSL for TMRs below 5 dB. The difference between SIR and SDR is that SIR does not penalize reverberation, implying that Sawada's system might reject more of the interference direct-path signal than MESSL while passing more reverberation, a hypothesis that is corroborated by the DERTM analysis. The DP-Oracle and Oracle results show that a gradual decline of SDR but not SIR as TMR decreases, possibly because of extra reverberation from the masking source creeping through the masks.

### B. Signal-to-noise ratio metrics

We also compare these systems using two other signal-to-noise ratio metric. The first, which we refer to as $\text{SNR}_{\text{ME}}$, was introduced by [37] and subsequently used in [38]. The second, which we refer to as $\text{SNR}_{\text{Li}}$, has been in wide use and is described by [13]. Both of these metrics penalize discarded desirable signal and passed undesirable signal. Unlike other metrics (e.g., [10], [11]), both of these metrics provide a single figure of merit. Typically we measure the signal-to-noise ratio improvement (SNRI), which is the difference between the SNR calculated for the specified mask and the SNR calculated for a mask of all 1s.

The $\text{SNR}_{\text{Li}}$ metric is calculated by measuring the ratio of the energy in the original target signal to the energy in the difference between the original target and the estimated target




(a) Source-to-distortion ratio (b) Source-to-interferer ratio (c) Source-to-artifact ratio

Fig. 6. BSS_EVAL evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio. Error bars show 1 standard error.

TABLE II
COMPARISON OF TWO SNR METRICS FOR EXTREME MASKS.

| Mask | $\text{SNR}_{\text{ME}}$ | $\text{SNR}_{\text{Li}}$ |
|---|---|---|
| All 0s | $-\infty$ | 0 |
| All 1s | mix SNR | mix SNR |
| Oracle | $\infty$ | $\infty$ |
| $1-$ Oracle | $-\infty$ | $<$ mix SNR |

signal. In the notation of Eq. (1), it is defined as

$$\text{SNR}_{\text{Li}} = 10\log_{10}\frac{\|s_j(t)\|^2}{\|s_j(t)-\hat{s}_j(t)\|^2}. \tag{7}$$

Because the energy in the original signal does not change, this metric has the convenient property of being relatively simple to optimize, notwithstanding issues of frame overlap [13]. It has some relatively strange properties in extreme conditions, however, which are shown in Tab. II and will be discussed shortly.

$\text{SNR}_{\text{ME}}$ is only applicable to mask-based source separation, and does not require a transformation back into the time domain from a spectrogram representation. It is similar to $\text{SNR}_{\text{Li}}$, but with a different numerator. Switching to a spectrogram-based notation, denote the mask $M(\omega,t)$, the desirable signal $S(\omega,t)$, and the sum of the undesirable signals $N(\omega,t)$. Omitting $(\omega,t)$ for readability, $\text{SNR}_{\text{ME}}$ is defined as

$$\text{SNR}_{\text{ME}} = 10\log_{10}\frac{\|MS\|^2}{\|(1-M)S+MN\|^2} \tag{8}$$

where $\|\cdot\|^2$ indicates the Frobenius norm of a spectrogram matrix, i.e. the sum of squares of all entries. The "signal" in this signal-to-noise ratio is the amount of target energy that survives the mask. The "noise" is a combination of energy rejected from the target source and energy accepted from the masker.

In this notation, Eq. (7) becomes

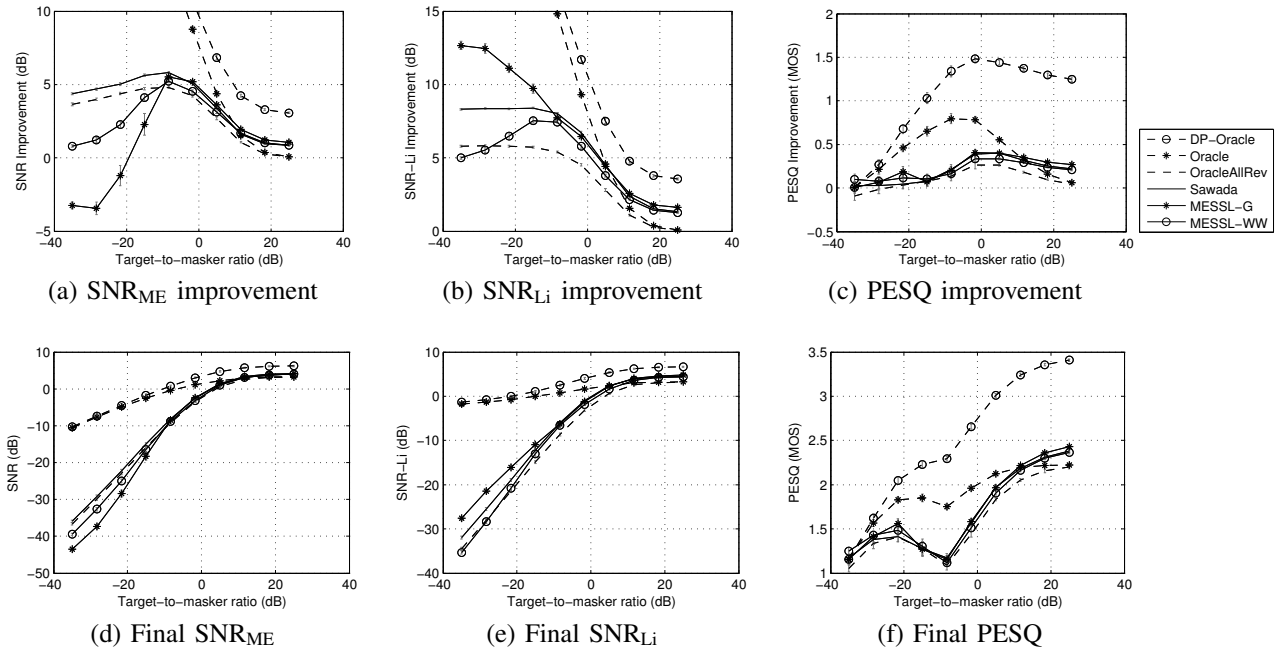
$$\begin{aligned}\text{SNR}_{\text{Li}} &= 10\log_{10}\frac{\|S\|^2}{\|S-M(S+N)\|^2}\\ &= 10\log_{10}\frac{\|S\|^2}{\|(1-M)S-MN\|^2}\end{aligned} \tag{9}$$

which is quite similar. Because the desirable and undesirable signals are generally uncorrelated (except when reverberation from a source in one is assigned to the other), there is little difference between the addition and subtraction in the denominators of Eqs. (8) and (9). The "signal" in $\text{SNR}_{\text{Li}}$ is the amount of energy in the original target signal. This number more or less serves as a constant baseline against which to compare the "noise", which is the difference between the reconstruction and the original target.

The main difference between these equations is in the numerators and certain examples highlight these differences, as shown in Tab. II. This table shows that $\text{SNR}_{\text{Li}}$ tends to estimate higher values than $\text{SNR}_{\text{ME}}$. In the case of a "separation" that passes no signal at all, the denominator in $\text{SNR}_{\text{Li}}$ is the entire original desirable signal, giving a ratio of 0 dB. $\text{SNR}_{\text{ME}}$, on the other hand, assigns this "separation" an SNR of $-\infty$. Similarly, when an oracle mask is designed to eliminate all of the target energy, $\text{SNR}_{\text{ME}}$ will assign a very large, negative SNR, while $\text{SNR}_{\text{Li}}$ will assign it some SNR less than the mixture SNR. It could be 0 dB or it could be less, but it depends on the *a priori* ratio between the desirable and undesirable signals. Although it is easier to optimize, we believe that these characteristics make $\text{SNR}_{\text{Li}}$ less appealing to use than our proposed $\text{SNR}_{\text{ME}}$ metric for evaluation purposes. These theoretical results are corroborated by Fig. 7, which will now be discussed.

The results of the $\text{SNR}_{\text{ME}}$ evaluation can be seen in Fig. 7(a) and (d). As in Fig. 6(a), the Oracle and DP-Oracle masks perform much better than the algorithmic masks at low TMRs, while the OracleAllRev performs very similarly to the algorithmic masks. Because the basic $\text{SNR}_{\text{ME}}$ contour shown in Fig. 7(d) is so closely related to the initial TMR, Fig. 7(a) shows the improvement each algorithm provides relative to that baseline. It can be seen again that for TMRs above $-10$ dB MESSL-G performs slightly better than the other algorithms, but below that TMR, Sawada's algorithm performs better. Note that both of these algorithms outperform OracleAllRev by approximately 1 dB in the regimes in which they perform best.

The results of the $\text{SNR}_{\text{Li}}$ evaluation can be seen in Fig. 7(b) and (e). These results are similar to those of (a) and (d), but with a number of significant differences that are more apparent in the improvement results, shown in subplots (a) and (b). The performance of the three ground truth masks under both of these metrics is actually quite similar. The performance of the algorithmic separations under these metrics is quite different, however. Most notably, MESSL-G, which other metrics have shown failed for TMRs below $-15$ dB, still appears to have improved the $\text{SNR}_{\text{Li}}$ by as much as 7 dB more than MESSL-WW, which did not fail as catastrophically as measured by




(a) $\text{SNR}_{\text{ME}}$ improvement (b) $\text{SNR}_{\text{Li}}$ improvement (c) PESQ improvement

(d) Final $\text{SNR}_{\text{ME}}$ (e) Final $\text{SNR}_{\text{Li}}$ (f) Final PESQ

Fig. 7. Objective evaluations of ground truth and algorithmic masking systems as a function of target-to-masker ratio. Error bars show 1 standard error.

other metrics. $\text{SNR}_{\text{Li}}$ does show MESSL-WW's divergence from Sawada and OracleAllRev for low TMR, but it also shows Sawada's algorithm as maintaining a large advantage over OracleAllRev that is not present using other metrics.

### C. PESQ

The final objective measure with which we evaluate these separations is the Perceptual Evaluation of Speech Quality (PESQ). This metric was originally designed to predict the *quality* of speech for telephony applications, not its intelligibility. In tests of its efficacy, it has been found to accurately predict not only the quality of speech isolated from mixtures by separation algorithms [5], but also ASR performance on speech isolated from mixtures [6].

PESQ compares a processed signal to its original. In the case of reverberant source separation, we compare the original, unmixed, anechoic signal to the signal estimated from the mixture by the source separation algorithms under evaluation. The crux of PESQ's computation is a calculation of the difference between the original signal and the processed signal after both have been transformed into a psychoacoustically-motivated representation [12, Sec. 10.5.3.3]. It uses a time-frequency representation with frequencies spaced according to the Bark scale and amplitudes measured in the sone scale. Differences between the signals are combined across time and frequency and positive and negative differences are treated differently, each being weighted by factors learned from a linear regression again human quality judgments.

The PESQ results on this experiment are shown in Fig. 7(c) and (f). The absolute PESQ score is shown in subplot (f) while subplot (c) shows the difference between this score and the PESQ score of the target source in the original mixture. Note that all of the algorithms' lines in subplot (f) are non-monotonic, with a secondary peak around $-20$ dB. This could be due to PESQ's behavior for noisy signals, because it was designed for speech. As the TMR is lowered in the mixture, the original target's amplitude is decreased and more of it is lost below the noise floor, leading to a noisier target signal and a less accurate estimate of quality.

As with the other metrics, the Oracle and DP-Oracle masks perform qualitatively differently from the other masks. With PESQ, however, they also perform qualitatively differently from each other. Most notably, the Oracle mask's PESQ improvement peaks at a TMR of around $-10$ dB and declines for TMRs higher or lower, while the DP-Oracle mask's PESQ improvement maintains the same level for higher TMRs. This is most likely due to distortions introduced by reverberation. Thus, the main cause of decreased PESQ scores for low TMRs is the masking source, while for high TMRs it is reverberation from the target source itself.

Of the algorithmic separations, MESSL-G has the highest PESQ scores across all TMRs. Because of MESSL-G's failure at low TMR, as shown by the other metrics, PESQ is probably not working properly for evaluating mixtures with low TMR. For TMRs above $-15$ dB, however, where it appears to be working, all of the algorithmic masks outperform the OracleAllRev mask. As can be seen for TMRs above 10 dB, this is most likely due to ancillary dereveberation that the algorithms perform. This is also corroborated by the fact that at those TMRs the algorithms also outperform the Oracle mask, which includes all of the target source's reverberation. Note, however, that in Sec. III, reverberation from the target source has a much smaller effect on ASR performance than PESQ estimates here.

## VII. Conclusions

This paper compared speech separation performance of a number of algorithmic and ground truth masks and examined

a number of metrics for evaluating the performance of such masks. Automatic speech recognition results should be the best predictor of intelligibility, but we did not find that ASR results for any of the masks under evaluation predicted important aspects of human intelligibility particularly well. Specifically, while the numerical ASR results for ground truth separations followed human intelligibility in anechoic conditions, they were much better than human performance in reverberation. Algorithmic separations, on the other hand, were also similar to human performance in anechoic conditions, but much worse than human performance in reverberation. These results suggest that the separation algorithms investigated in this paper cannot reject reverberant energy as well as humans can.

The use of ground truth masks that treat reverberation differently from direct-path signals supports the idea that poor performance of algorithmic separations comes from a failure to discount reverberant energy. This conclusion is corroborated by the DERTM analysis, which examines the effect of masks on various component signals from the mixture. Both of these analyses imply that while the algorithmic masks successfully suppress the direct-path signal coming from a masking source, they do not sufficiently suppress its reverberation.

In examining other metrics that are commonly used in the literature, we found that the SDR metric from the BSS_EVAL toolbox agrees well with the ASR results, as does our proposed $\mathrm{SNR}_{\mathrm{ME}}$ metric. Both of these metrics correctly indicate that MESSL's separation fails for TMRs below $-15$ dB, while the algorithm of Sawada et al. does not. Both metrics are better predictors of ASR performance than the SIR and SAR metrics from the BSS_EVAL toolbox and the $\mathrm{SNR}_{\mathrm{Li}}$ metric, which do not indicate the failure of MESSL for very low TMRs.

The failure of MESSL for very low TMRs is most likely due to the failure of the localization algorithm used to initialize it. The fact that Sawada's algorithm does not depend on such an initialization makes it more robust at these very low TMRs. Although it was not investigated in this paper, it should be possible to use a hybrid model that initializes MESSL from the mask generated by Sawada's algorithm when it detects that its localization has failed, which could yield performance equal to or better than both of the original algorithms.

Overall, better models of reverberation or other diffuse noise cancellation mechanisms seem necessary to improve ASR performance significantly in reverberant mixtures. We believe that such models would also improve the intelligibility of separations, although this result is not directly addressed by the experiments of this paper.

## ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under Grant No. IIS-0535168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

## REFERENCES

[1] D. P. W. Ellis, "Evaluating speech separation systems," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Kluwer, 2004, ch. 20, pp. 295–304.

[2] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acous. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[3] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *J. Acous. Soc. Amer.*, vol. 124, no. 4, pp. 2303–2307, 2008.

[4] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University department of Electrical Engineering, 1985.

[5] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Tr. Audio, Speech, and Language Proc.*, vol. 16, no. 1, pp. 229–238, 2008.

[6] L. Di Persia, D. Milone, H. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, October 2008.

[7] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, October 1995.

[8] C. M. Chernick, S. Leigh, K. L. Mills, and R. Toense, "Testing the ability of speech recognizers to measure the effectiveness of encoding algorithms for digital speech transmission," in *IEEE Military Communications Conference*, vol. 2, 1999, pp. 1468–1472.

[9] W. Jiang and H. Schulzrinne, "Speech recognition performance as an effective perceived quality predictor," in *IEEE International Workshop on Quality of Service*, 2002, pp. 269–275.

[10] G. Hu and D. Wang, "Monaural speech separation," in *Adv. Neur. Inf. Proc. Sys.*, vol. 15, 2003, pp. 1221–1228.

[11] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[12] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC, June 2007.

[13] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Sp. Comm.*, vol. 51, no. 3, pp. 230–239, March 2009.

[14] S. Bressler and B. G. Shinn-Cunningham, "Effects of pitch and spatial separation on selective attention in anechoic and reverberant environments," *J. Acous. Soc. Amer.*, vol. 123, no. 5, p. 2978, 2008.

[15] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acous. Soc. Amer.*, vol. 109, no. 3, pp. 1101–1109, 2001.

[16] D. S. Brungart, B. D. Simpson, and R. L. Freyman, "Precedence-based speech segregation in a virtual auditory environment," *J. Acous. Soc. Amer.*, vol. 118, no. 5, pp. 3241–3251, 2005.

[17] G. Kidd, T. L. Arbogast, C. R. Mason, and F. J. Gallun, "The advantage of knowing where to listen," *J. Acous. Soc. Amer.*, vol. 118, no. 6, pp. 3804–3815, 2005.

[18] C. R. Mason, T. L. Rohtla, and P. S. Deliwala, "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," *J. Acous. Soc. Amer.*, vol. 104, no. 1, pp. 422–431, 1998.

[19] R. L. Freyman, U. Balakrishnan, and K. S. Helfer, "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," *J. Acous. Soc. Amer.*, vol. 115, no. 5, pp. 2246–2256, 2004.

[20] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, May 2008.

[21] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acous. Soc. Amer.*, vol. 117, no. 5, pp. 3100–3115, 2005.

[22] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Tr. Audio, Speech, and Language Proc.*, 2009, in press.

[23] A. Jourjine, S. Rickard, and O. Yılmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 5, 2000, pp. 2985–2988.

[24] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 139–142.

[25] J. Mouba and S. Marchand, "A source localization / separation / respatialization system based on unsupervised classification of interaural cues," in *Proc. Int. Conf. on Digital Audio Effects*, 2006, pp. 233–238.

[26] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Sp. Comm.*, vol. 43, no. 4, pp. 275–296, September 2004.

[27] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Sp. Comm.*, vol. 34, no. 3, pp. 267–285, June 2001.
[28] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 82–85.
[29] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, vol. 2, 1990, pp. 845–848.
[30] R. J. Weiss and D. P. W. Ellis, "Monaural speech separation using source-adapted models," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 2007, pp. 114–117.
[31] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech & Language*, January 2009.
[32] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Sp. Comm.*, vol. 43, no. 4, pp. 379–393, September 2004.
[33] H.-M. Park and R. M. Stern, "Missing feature speech recognition using dereverberation and echo suppression in reverberant environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 4, 2007, pp. 381–384.
[34] S. A. Gelfand and S. Silman, "Effects of small room reverberation upon the recognition of some consonant features," *J. Acous. Soc. Amer.*, vol. 66, no. 1, pp. 22–29, 1979.
[35] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Independent Component Analysis and Signal Separation*, 2009, pp. 734–741.
[36] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Tr. Audio, Speech, and Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.
[37] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neur. Inf. Proc. Sys.*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 953–960.
[38] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 275–278.

**Michael I. Mandel** received the B.S. degree in Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, in 2004 and the M.S. and Ph.D. degrees in Electrical Engineering from Columbia University, New York, NY, in 2006 and 2010 in Prof. D. Ellis' Laboratory for the Recognition and Organization of Speech and Audio.

He is currently a postdoctoral researcher in the Machine Learning laboratory at the Université de Montréal in Montréal, Québec. He has published on sound source separation, music similarity, and music recommendation. His research uses machine learning to model sound perception and understanding.

**Scott Bressler** received the B.A. degree in Biology in 1994 and the M.S. degree in Biomedical Engineer in 2007 from Boston University. He is currently a Research Associate in the Auditory Neuroscience Laboratory at the Department of Cognitive and Neural Systems at Boston University.

**Barbara Shinn-Cunningham** (Brown University, Sc.B.; MIT, M.S. and Ph.D.) is a Professor of Cognitive and Neural Systems, Biomedical Engineering, and the Program in Neuroscience at Boston University, where she has numerous administrative responsibilities (Director of Graduate Studies, member of the Executive Committee of the BU Center for Neuroscience, co-Director of the CELEST Science of Learning Center, etc.). She chairs the AUD study section for NIH, is an Associate Editor for the Journal of the Associate for Research in Otolaryngology, and holds leadership positions in numerous professional societies, including the Association for Research in Otolaryngology and the Acoustical Society of America. Her research on attention, auditory perception, and spatial hearing has lead to fellowships from the Alfred P. Sloan Foundation, the Whitaker Foundation, and the National Security Science and Engineering Faculty Fellows program. She is a lifetime National Associate of the National Research Council of the National Academies of Science and a Fellow of the Acoustical Society of America.

**Daniel P. W. Ellis** received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor in the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound. He worked at MIT, where he was a Research Assistant at the Media Lab, and he spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA.