Chapter 6

Evaluation

This chapter examines the performance of the algorithms from chapter 5 on a new 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 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. We also find that MESSL breaks down at very low target-to-masker ratios because of a failure of its localization initialization and that this breakdown is not detected by PESQ or certain signal-to-noise ratio metrics that are typically used in similar evaluations.

6.1 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 overdetermined mixture (i.e. a mixture with more microphones than sources), perfect reconstruction of each source should be attainable. When separating sources from an underdetermined mixture, 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.

Depending on the intended application of a source separator, it usually makes sense to compare the original and estimated signals in a representation that is more relevant to the application than the raw waveform (Ellis, 2004). 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. This chapter focuses on measuring the performance improvement that source separators give to ASR systems. It does not address the intelligibility of these separations: no subjects ever listened to the processed signals.

Because human speech recognition is still well beyond 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 (Brungart et al., 2006; Wang et al., 2008), 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. Weintraub, 1985).

ASR and intelligibility both provide a somewhat rough 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 PESQ and SDR metrics described in chapter 5, can be informative, they have not been shown decisively to correspond well with intelligibility of signals for human subjects.

The Perceptual Evaluation of Speech Quality (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 (Hu and Loizou, 2008). It has also been used to predict ASR performance (Di Persia et al., 2008). Quality is a property of speech that corresponds to its realism and naturalness, characteristics that are not necessary for intelligibility. Wang et al. (2008) and Shannon et al. (1995) 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 (Chernick et al., 1999; Jiang and Schulzrinne, 2002).

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. Such metrics include the SDR used in chapter 5. As measures of intelligibility, these metrics are flawed, since signal energy does not correlate consistently with intelligibility (Ellis, 2004). For example, low speech frequencies tend to be relatively intense, independent of spectral and temporal 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., Hu and Wang, 2003; Yilmaz and Rickard, 2004), complicating comparisons of different algorithms and making it difficult to determine how a given algorithm is failing.

After describing the experiment that this chapter is based on in section 6.2, section 6.3 evaluates the performance of automatic speech recognition on the outputs of each of the separation algorithms described in chapter 5. Section 6.4 then examines the effect of each of these algorithms on various components of the target and masker signals using a novel technique we call DERTM, which distinguishes between different types of errors, enabling better evaluation of when a system fails. Finally section 6.5 compares these analyses with metrics that are commonly used in the speech separation literature including PESQ, SNR_{Li} (our designation), and BSS_EVAL, and discusses an energy-based metric that we introduced.

6.2 Shared experimental details

The experiment in this chapter is based on the human spatial hearing experiment of Bressler and Shinn-Cunningham (2008), but only includes a subset of the conditions tested in that paper. The next section describes the original experiment, and the subsequent section describes the experiment that we carried out based on it.

6.2.1 Details of original experiment

The experiment of Bressler and Shinn-Cunningham (2008) examined the intelligibility of speech in reverberant and anechoic conditions while varying the cues that the 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 normal, 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° 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 Brungart, 2001; Brungart et al., 2005; Kidd et al., 2005a; Jr et al., 1998; Freyman et al., 2004), essentially solving the problem of source selection (see Shinn-Cunningham, 2008). Instead, listeners should be 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. Firstly, masking was maximal when intonation and spatial cues were the same between the two utterances. In reverberation, where segregation cues are weakened, any difference between the two sources provided a release from masking. A single difference between the sources increased intelligibility and a second difference increased it even more. 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 between configurations.

Generally, in reverberant conditions, the curve of intelligibility vs target-to-masker ratio (TMR) (see figures 6.2 and 6.3 for these curves) was shifted towards higher TMRs and had a more shallow slope than in anechoic conditions. 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).

6.2.2 Details of current experiment

The experiments performed in this chapter used the same utterances and the same impulse responses as Bressler and Shinn-Cunningham (2008). The impulse responses were recorded in a classroom with the source 1.2 m from a KEMAR dummy head and have a reverberation time of approximately 550 ms and a direct-to-reverberant ratio of approximately 10 dB. One binaural impulse response was recorded at each of a number of azimuths; however, we only use the two recorded at 0° and 90°. Because localization is central to separation in all of the systems under evaluation, we did not evaluate them on the collocated conditions. Unlike in the experiment described above, 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

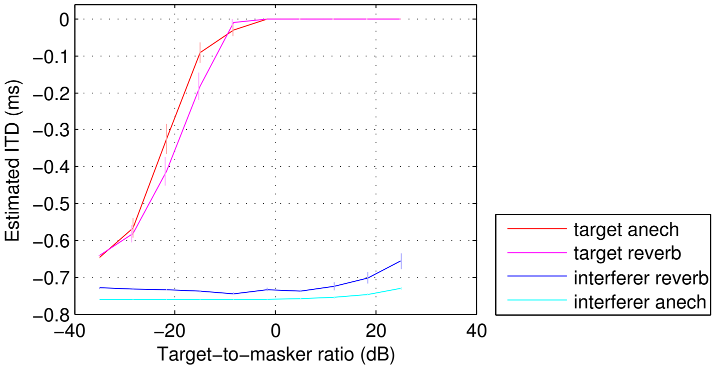


Figure 6.1: Localization as a function of target-to-masker ratio. Results are shown for anechoic and reverberant conditions. The sources are located at 0 ms and 0.76 ms. Error bars show 1 standard error.

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.

All six of the oracle separation algorithms described in section 3.2 were tested. While all of the ground truth masks are constructed based on the ratio of desirable energy to undesirable energy at each time-frequency point, these masks differ in their classification of these four types of energy as desirable or not. See table 3.1 for an indication of the classification of these signals in each mask. Note that for the purposes of constructing these masks, early echoes were considered to be part of the reverberation, as they did not make an appreciable difference on performance.

6.2.3 Algorithmic separations

The algorithmic separation systems used in this experiment are the same as those used in section 5.4.5, except that TRINICON was not evaluated. While all of them are compared in section 6.3, 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. MESSL was run with the $\Theta_{\Omega\Omega}$ parameters, with and without the garbage source and ILD prior. We again denote the system using the garbage source and ILD prior as the Θ_G parameters.

The localization performance of MESSL is shown in figure 6.1 as a function of target-to-masker ratio (TMR). In the anechoic mixtures the localization performance is more accurate, especially for the source at 90° (the masker), and accurate over a wider range of TMRs than in the reverberant mixtures. For reverberant mixtures, the localization begins to fail when one source is more than 15 dB louder than the other. Note that the failure is not symmetric in TMR. This is because the masker was always longer in duration than the target, so that even when the target was much louder, the masker was still the only source audible for a short time. Even this brief glimpse was sufficient to allow a more robust localization.

6.3 Automatic speech recognition vs intelligibility

As described in sections 2.4.3 and 2.4.4, 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 Weintraub (1985). 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.

6.3.1 ASR setup

While many papers describe a single recognizer with which a number of different separators are tested (Seltzer et al., 2004; Park and Stern, 2007), 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 as described in section 2.4.4 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 MFCCs, MFCC deltas, and MFCC double-deltas.

Three different recognizers were trained for each algorithm, one on only anechoic separations, one on only reverberant separations, and one on both, which was subsequently evaluated on both anechoic and reverberant separations. As described in section 6.2.2, the three datasets used separate target utterances, so there was no opportunity to memorize the target utterances. They did use the same maskers, but because these were never recognized, they should not affect the results. In total, there were 3200 digits in each of the tuning and test sets and 2560 digits in the training set. The mixtures were randomized in the test set at different TMRs to provide as many different contexts of each digit as possible for the recognizer. 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, the system's transcription and the base transcription were aligned using the minimum edit distance with equal weights. The fraction of digits that were correctly identified (COR) was used as the measure of performance. Note that this is not the same as the complement of word error rate (ACC), which uses the number of digits correctly identified minus the number of insertions, according to

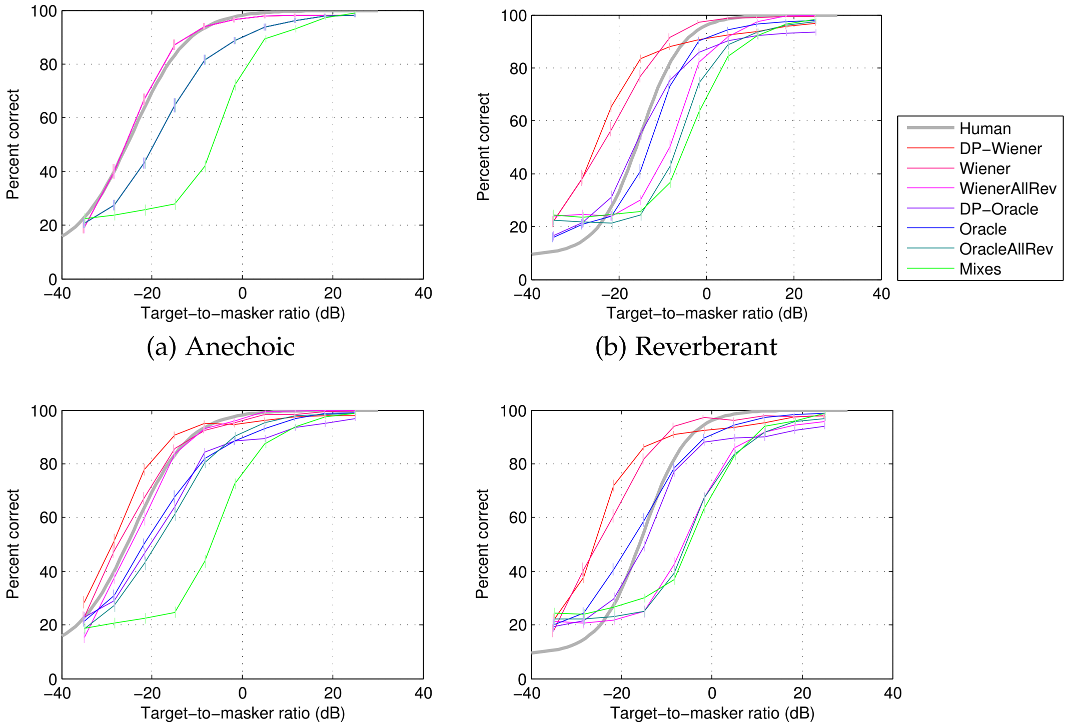
$$\text{COR} = \frac{C}{N} \qquad \text{ACC} = \frac{C - I}{N} \tag{6.1}$$

where C is the number of correct digits, I is the number of insertions, and N is the total number of digits. With a fixed sentence length of five digits, every insertion must be accompanied by a deletion. For example, if the system transcribed "11234" when the correct digit string was "12345", it would have inserted an extra 1 and deleted the 5. The scoring in this case would be $C = 4$, $I = 1$, $N = 5$, so COR $= 80\%$ while ACC $= 60\%$. Thus, the ACC metric penalizes the insertion twice. One problem with the COR metric is that in correcting for insertions it may be too forgiving when an algorithm is randomly guessing. In particular, this is evident in the performance of the ASR systems shown in figures 6.2 and 6.3, which seem to asymptote to 20% correct at very low TMRs instead of 9% one would expect if each digit was an independent random guess.

6.3.2 Results

The results of the ASR experiment for ground truth separations can be seen in figure 6.2. Human performance on this task 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 green curve. In figure 6.2(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). In figure 6.2(c) the addition of reverberant mixtures to the training data differentiates the masks, and it can be seen that the systems including less reverberation in training perform slightly better in testing here. For both training conditions, the recognizers perform at close to human levels, particularly the Wiener masks, and significantly better than recognizers trained and tested on the mixtures.

In reverberation, the results are slightly different. Figure 6.2(b) shows the results of the ASR systems trained only on reverberant separations. 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



(a) Anechoic

(b) Reverberant

(c) Anechoic, trained on both

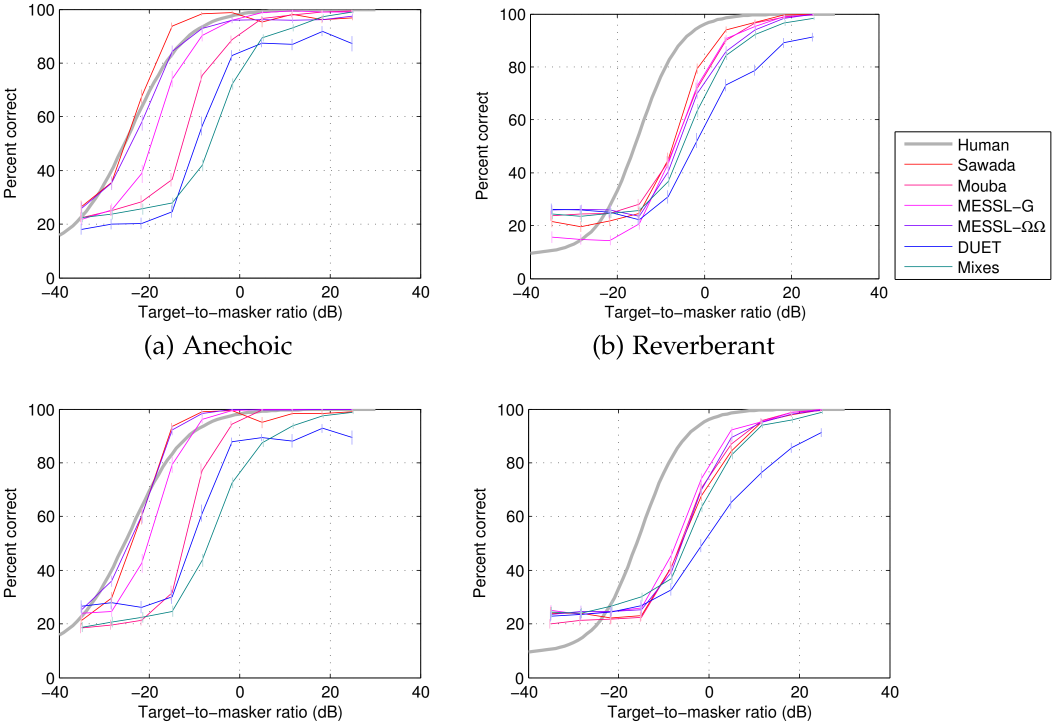
(d) Reverberant, trained on both

Figure 6.2: Automatic speech recognition performance as a function of target-to-masker after applying ground truth separation masks. Top row recognizers were trained only on matched anechoic or reverberant data, bottom row recognizers were trained on both anechoic and reverberant data, and tested on each separately. Note that in (a) many of the lines lie exactly on top of each other. Error bars show 1 standard error.

unseparated mixtures. The addition of anechoic training data in figure 6.2(d) does not change these results much, although the WienerAllRev recognizer is no longer able to outperform the mixtures.

The results of the ASR experiment for algorithmic separations can be seen in figure 6.3. Once again, human performance is indicated by the thicker gray line and the performance on unprocessed mixtures is indicated by the olive line. Figure 6.3(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. Results are similar when trained on both anechoic and reverberant separations, as seen in figure 6.3(c), although DUET's performance increases and Sawada's performance decreases when trained on both.

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



(a) Anechoic

(b) Reverberant

(c) Anechoic, trained on both

(d) Reverberant, trained on both

Figure 6.3: Automatic speech recognition performance as a function of target-to-masker ratio after applying algorithmic separation masks. Top row recognizers were trained only on matched anechoic or reverberant data, bottom row recognizers were trained on both anechoic and reverberant data, and tested on each separately. Error bars show 1 standard error.

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. Section 6.4 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. Section 6.5 examines objective measures of speech intelligibility and quality.

6.4 Six sub-signal attenuations

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. Because

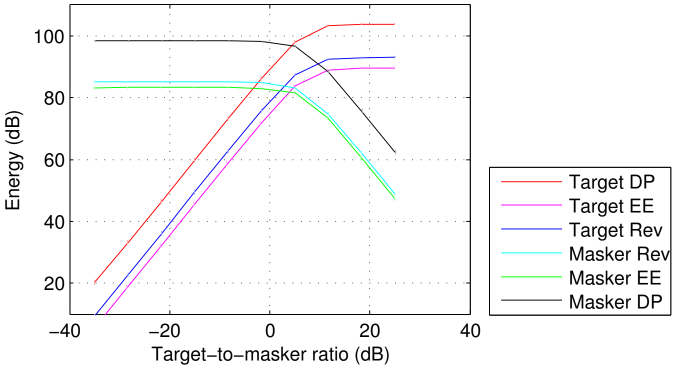


Figure 6.4: Initial levels of the six different types of energy: target direct-path, target reverberation, masker direct-path, masker reverberation. Error bars show 1 standard error.

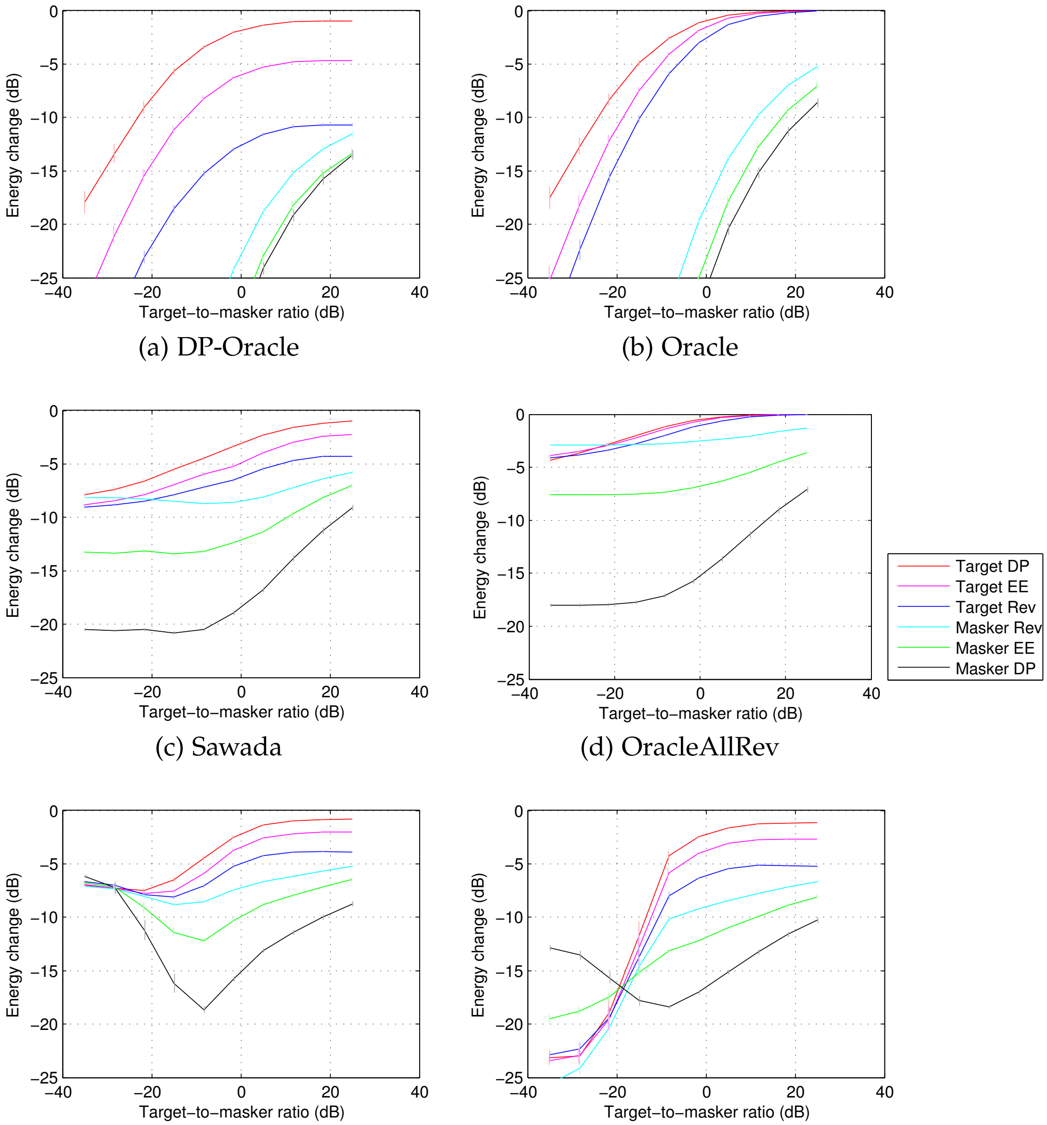
of the similarity of the binary and Wiener ground truth systems, we will only examine the performance of the three binary ground truth separators. Because of their better performance, we will only compare Sawada, MESSL-G, and MESSL-WW in this and subsequent sections. And because it is a more difficult and realistic task, we will also only examine the performance of these systems on the reverberant mixtures.

As mentioned in the discussion of ground truth masks, 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 initial levels of each of these signals in the mixtures can be seen in figure 6.4. The piecewise linear trend of the graph is due to the mixtures' being scaled to a constant level after combining the sources at the specified TMR. Note that the direct-to-reverberant ratio can be read from this graph, and is approximately 10 dB for both the target and the masker.

Because all of the masks under evaluation use values between 0 and 1, they are all subtractive, meaning that they only delete energy, they do not add it. Furthermore, as described in section 2.1, we have access to the original versions of these six signals for every mixture and can easily measure the attenuation of each of them by a given mask[1]. It is informative to compare this analysis for each algorithm to the ASR results in section 6.3.

The results of this analysis are shown in figure 6.5 for each algorithm relative to the initial levels in figure 6.4. 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 and all of the energy in a given point, whether desirable or not, must be treated identically.

[1] Note that this does not account for the potential "transformation" of masker energy into target energy, as in (Wang et al., 2008), although such a transformation is probably less likely to happen in pure speech-speech mixtures than in the speech-noise mixtures those authors examined



(a) DP-Oracle

(b) Oracle

(c) Sawada

(d) OracleAllRev

(e) MESSL using $\Theta_{\Omega\Omega}$

(f) MESSL using Θ_G

Figure 6.5: DERTM analysis of masking separations. Masking 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 indicate 1 standard error.

The DP-Oracle mask is able to attenuate the other signals by a large 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 (Gelfand and Silman, 1979). 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 using $\Theta_{\Omega\Omega}$ performs similarly for high TMRs, but clearly begins to break down for TMRs below -10 dB, in agreement with the localization data shown in figure 6.1. 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 chapter. 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 a worse job of rejecting reverberation from either source and in rejecting the direct-path of the masking source than Sawada's algorithm.

MESSL using the Θ_G parameters also fails for very low TMRs, but it does so in a different manner than when using $\Theta_{\Omega\Omega}$. 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. This is perhaps a better failure mode than MESSL using $\Theta_{\Omega\Omega}$ because the masker direct-path is attenuated by an additional 8 dB. 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. Any energy from the target that is present is absorbed in the garbage source along with reverberation from the masking source. The source selected with the closest localization to the target only has a small fraction of the masker energy, causing its 14 dB attenuation. All of the other signals are absorbed in the garbage source, causing their attenuation by 20 dB.

6.5 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.

6.5.1 BSS_EVAL metrics

As described in section 5.4.1.2, tools from the Signal Separation Evaluation Campaign (Vincent et al., 2009) and the BSS_EVAL toolbox (Vincent et al., 2006) can be used to 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 $P(x, \{y_i\}, \tau_{\max})$ to be the projection of signal x onto versions of the signals $\{y_i\}$ shifted by every integral number of samples up to $\tau_{\max}$. Thus for signals T samples long, this projection uses a matrix that is $T \times T$ and rank $I\tau_{\max}$ 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

$$s_{\text{target}}(t) = P(\hat{s}_j, s_j, \tau_{\max}) \tag{6.2}$$

$$e_{\text{interf}}(t) = P(\hat{s}_j, \{s_i\}, \tau_{\max}) - P(\hat{s}_j, s_j, \tau_{\max}) \tag{6.3}$$

$$e_{\text{artif}}(t) = \hat{s}_j - P(\hat{s}_j, \{s_i\}, \tau_{\max}) \tag{6.4}$$

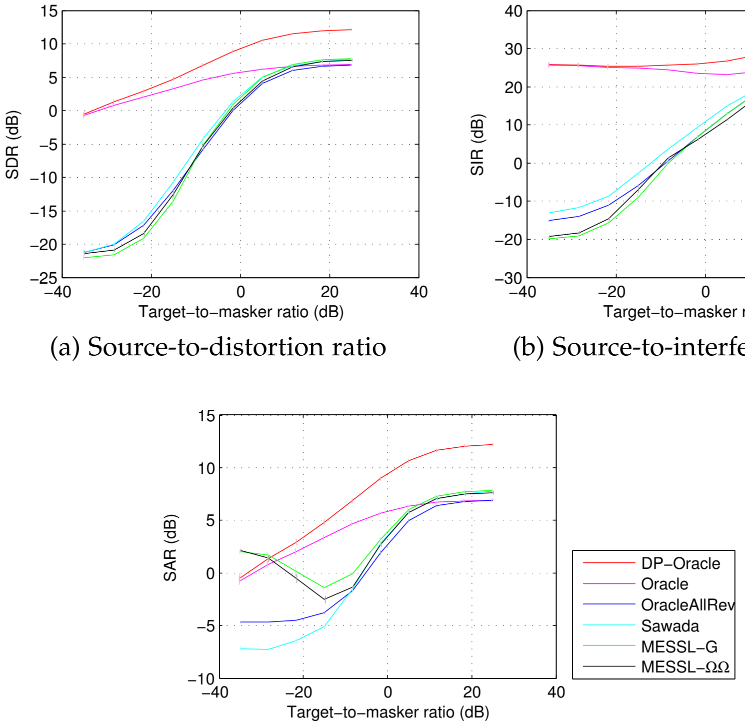
The three ratios of interest are defined as

$$\text{SDR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2} \tag{6.5}$$

$$\text{SIR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{6.6}$$

$$\text{SAR} = 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2}. \tag{6.7}$$

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_{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



(a) Source-to-distortion ratio

(b) Source-to-interferer ratio

(c) Source-to-artifact ratio

Figure 6.6: BSS_EVAL evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio. Error bars show 1 standard error.

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.

Figure 6.6 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 section 6.4. 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

Table 6.1: Comparison of two SNR metrics for extreme masks.

Mask	SNR_{ME}	SNR_{Li}
All 0s	$-\infty$	0
All 1s	mix SNR	mix SNR
Oracle	∞	∞
$1 -$ Oracle	$-\infty$	$<$ mix SNR

low TMRs.

Under the SIR metric, as seen in figure 6.6(b), Sawada's algorithm performs significantly better than MESSL for TMRs below 5 dB. The difference between SIR and SDR is that reverberation is not penalized by SIR, 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 the gradual decline of SDR but not SIR as TMR decreases could be a result of extra reverberation from the masking source creeping through the masks.

6.5.2 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 SNR_{ME}, was introduced by Mandel et al. (2007) and subsequently used in (Mandel and Ellis, 2007). The second, which we refer to as SNR_{Li}, has been in wide use and is described by (Li and Wang, 2009). Both of these metrics penalize discarded desirable signal and passed undesirable signal. Unlike other metrics (e.g. Hu and Wang, 2003; Yilmaz and Rickard, 2004), 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 SNR_{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 signal. In the notation of equation (6.2), it is defined as

$$\text{SNR}_{\text{Li}} = 10 \log_{10} \frac{\|s_j(t)\|^2}{\|s_j(t) - \hat{s}_j(t)\|^2}. \tag{6.8}$$

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 (Li and Wang, 2009). It has some relatively strange properties in extreme conditions, however, which are shown in table 6.1 and will be discussed shortly.

SNR_{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 SNR_{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 (ω, t) for readability, SNR_{ME} is defined as

$$\text{SNR}_{\text{ME}} = 10 \log_{10} \frac{\|MS\|^2}{\|(1-M)S + MN\|^2} \tag{6.9}$$

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, equation (6.8) becomes

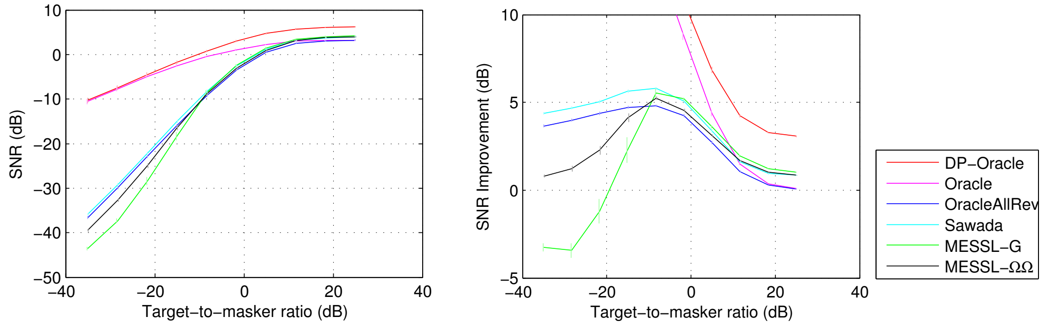
$$\text{SNR}_{\text{Li}} = 10 \log_{10} \frac{\|S\|^2}{\|S - M(S+N)\|^2} = 10 \log_{10} \frac{\|S\|^2}{\|(1-M)S - MN\|^2} \tag{6.10}$$

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 equations (6.10) and (6.9). The “signal” in SNR_{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 table 6.1. This table shows that SNR_{Li} tends to estimate higher values than SNR_{ME}. In the case of a “separation” that passes no signal at all, the denominator in SNR_{Li} is the entire original desirable signal, giving a ratio of 0 dB. SNR_{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, SNR_{ME} will assign a very large, negative SNR, while SNR_{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 SNR_{Li} less appealing to use than our proposed SNR_{ME} metric for evaluation purposes. These theoretical results are corroborated by figures 6.7 and 6.8, which will now be discussed.

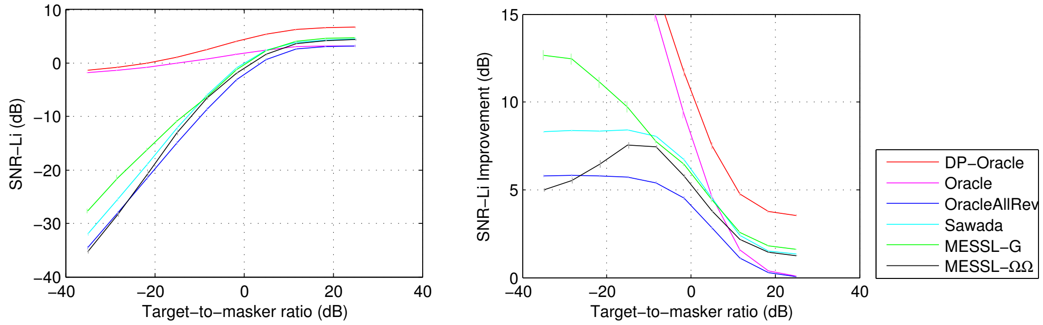
The results of the SNR_{ME} evaluation can be seen in figure 6.7. As in figure 6.6(a), the Oracle and DP-Oracle masks perform much better than the algorithmic masks at low TMRs, while the OracleAllRev performs very similarly to the algorithmic masks. Because the basic SNR_{ME} contour shown in figure 6.7(a) is so closely related to the initial TMR, figure 6.7(b) 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 SNR_{Li} evaluation can be seen in figure 6.8. These results are similar to those of figure 6.7 but with a number of significant differences that are more apparent in the improvement results, shown in subplot (b) of each figure. The performance of the three ground truth masks under both of these metrics is actually quite similar. The performance



(a) Final $\mathrm{SNR}_{\mathrm{ME}}$ (b) $\mathrm{SNR}_{\mathrm{ME}}$ improvement

Figure 6.7: SNR evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio. Error bars show 1 standard error.



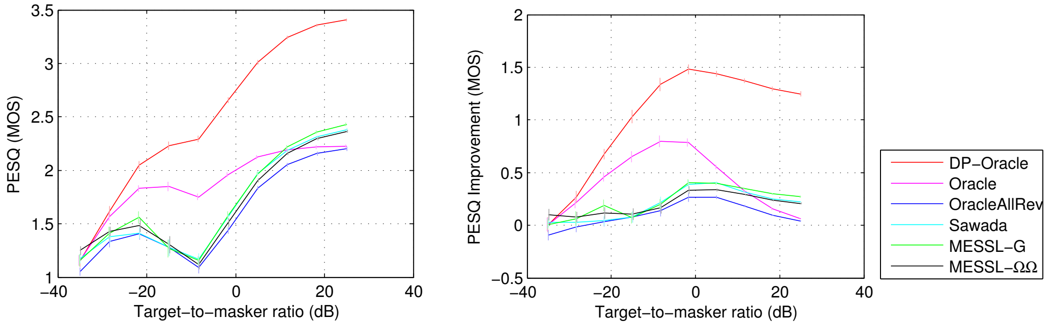
(a) Final $\mathrm{SNR}_{\mathrm{Li}}$ (b) $\mathrm{SNR}_{\mathrm{Li}}$ improvement

Figure 6.8: $\mathrm{SNR}_{\mathrm{Li}}$ evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio. Error bars show 1 standard error.

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 $\mathrm{SNR}_{\mathrm{Li}}$ by as much as 7 dB more than MESSL-WW, which did not fail as catastrophically as measured by other metrics. $\mathrm{SNR}_{\mathrm{Li}}$ 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.

6.5.3 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



(a) Final PESQ (b) PESQ improvement

Figure 6.9: PESQ evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio. Error bars show 1 standard error.

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 (Hu and Loizou, 2008), but also ASR performance on speech isolated from mixtures (Di Persia et al., 2008).

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 (Loizou, 2007, 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 figure 6.9. The absolute PESQ score is shown in subplot (a) while subplot (b) 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 (b) 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 section 6.3, reverberation from the target source has a much smaller effect on ASR performance than PESQ estimates here.

6.6 Summary

This chapter 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. We did not find, however, that any of the masks under evaluation predicted important aspects of human intelligibility results 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 separation algorithms 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 chapter, 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 the original Sawada and MESSL 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.