Technical Report OSU-CISRC-2/14-TR05
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2014**
File: **TR05.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

On Training Targets For Supervised Speech Separation

Yuxuan Wang
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
wangyuxu@cse.ohio-state.edu

Arun Narayanan
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
narayaar@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract – Formulation of speech separation as a supervised learning problem has shown considerable promise. In its simplest form, a supervised learning algorithm, typically a deep neural network, is trained to learn a mapping from noisy features to a time-frequency representation of the target of interest. Traditionally, the ideal binary mask (IBM) is used as the target because of its simplicity and large speech intelligibility gains. The supervised learning framework, however, is not restricted to the use of binary targets. In this study, we evaluate and compare separation results by using different training targets, including the IBM, the target binary mask, the ideal ratio mask (IRM), the short-time Fourier transform spectral magnitude and its corresponding mask (FFT-MASK), and the Gammatone frequency power spectrum. Our results in various test conditions reveal that the two ratio mask targets, the IRM and the FFT-MASK, outperform the other targets in terms of objective intelligibility and quality metrics. In addition, we find that masking based targets, in general, are significantly better than spectral envelope based targets. We also present comparisons with recent methods

in non-negative matrix factorization and speech enhancement, which show clear performance advantages of supervised speech separation.

Index Terms – Speech separation, deep neural networks, training targets, supervised learning.

1 Introduction

Speech separation, which is the task of separating speech from a noisy mixture, has major applications, such as robust automatic speech recognition (ASR), hearing aids design, and mobile speech communication. Monaural speech separation (i.e., speech separation from single-microphone recordings) is perhaps most desirable from the application standpoint. Compared to multi-microphone solutions, a monaural system is less sensitive to room reverberation and spatial source configuration. On the other hand, monaural separation is a severely underdetermined figure-ground separation problem. This study focuses on monaural separation.

Monaural speech separation has been widely studied in the speech and signal processing community for decades. From the signal processing viewpoint, many methods have been proposed to estimate the ideal Wiener filter, which is the optimal filter to recover clean speech in the minimum mean squared error (MMSE) sense [20]. A popular alternative to Wiener filtering is statistical model-based methods [20], [12], which infer speech spectral coefficients given noisy observations under prior distribution assumptions for speech and noise. Signal processing based methods usually work reasonably well in relatively high signal-to-noise ratio (SNR) conditions. However, they are generally less effective in low SNR and non-stationary noise conditions [20].

In contrast to signal processing based methods, model-based methods build models of speech and/or noise using premixed signals and show promising results in challenging conditions. For example, techniques in [24], [13] build probabilistic interaction models between different sources based on learned priors, and show significant performance gain in low SNR conditions. Another line of work is non-negative matrix factorization (NMF) [29], where noisy observations are modeled as weighted sums of non-negative source bases. These model-based methods work well if underlying assumptions are met. However, in our experience, these methods do not generalize well to unseen noisy conditions and are mostly effective for structured interference, e.g. music or a competing speaker. Moreover, these methods often require expensive inference, making them hard to use in real-world speech applications.

Recently, we have formulated monaural speech separation as a supervised learning problem, which is a data driven approach. In the simplest form, acoustic features are extracted from noisy mixtures to train a supervised learning algorithm, e.g. a deep neural network (DNN) [34]. In many previous studies (e.g. [15], [16], [9]), the training target (or the learning signal) is set to the ideal binary mask (IBM), which is a binary mask constructed from premixed speech and noise signals (see Section 3.1 for definition). This simplifies speech separation to a binary classification problem, a well studied machine learning task. Furthermore, IBM processing has been shown to yield large speech intelligibility improvements even in extremely low SNR conditions [1], [2], [18], [31]. Supervised speech separation aiming to estimate the IBM has shown a lot of promise. Notably, this approach has provided the first demonstration of improved speech intelligibility in noise for both normal hearing [16] and hearing impaired

listeners [11]. Supervised speech separation has also been shown to generalize well given sufficient training data [34], [36]. In addition, the system operates in a frame-by-frame fashion and inference is fast (only involving matrix operations), making it amenable to real-time implementation.

A suitable training target is important for supervised learning. On the one hand, one should use a target that can substantially improve speech perception in noise. On the other hand, the mapping from features to the target of interest should be amenable for training, say in terms of optimization difficulty [8]. Although the IBM is the optimal binary mask, it may not necessarily be the best target for training and prediction. Separation using binary gains typically produces residual musical noise. Other ideal targets are possible and can potentially improve speech intelligibility and/or quality, such as the target binary mask (TBM) [1], [17], the ideal ratio mask (IRM), the short-time Fourier transform (STFT) spectral magnitude, and the Gammatone frequency power spectrum. We note that some of them have been used in our preliminary work [21], [35], [10]. However, what training targets are appropriate for supervised speech separation remains unclear. This is clearly an important question with potentially important implications for separation performance. This paper addresses this question systematically, including a study of new training targets. In addition, we compare supervised separation with NMF and speech enhancement methods.

The rest of the paper is organized as follows. We first describe the DNN based supervised speech separation framework and the various training targets that we evaluate in the next two sections. Experimental settings and evaluation and comparison results are presented in Section 4 and 5, respectively. Discussions and conclusions are provided in Section 6.

2 Supervised Speech Separation

Speech separation can be interpreted as the process that maps a noisy signal to a separated signal with improved intelligibility and/or perceptual quality[1]. Without considering the impact of phase, this is often treated as the estimation of clean speech magnitude or some ideal mask. Supervised speech separation formulates this as a supervised learning problem such that the mapping is explicitly learned from data. Acoustic features are extracted from a mixture, which, along with the corresponding desired outputs are fed into a learning machine for training. New noisy mixtures are separated by passing estimated outputs and mixture phase into a resynthesizer.

To focus our study on learning targets, we use a fixed set of complementary features [32] throughout the experiments. The feature set includes amplitude modulation spectrogram (AMS), relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP), mel-frequency cepstral coefficients (MFCC), and 64-channel Gammatone filterbank power spectra (GF). All these features are extracted at the frame level and are concatenated with

[1]Depending on the application, the desired output of the mapping does not have to be clean speech.

the corresponding delta features. We also employ an auto-regressive moving average (ARMA) filter [3] to smooth temporal trajectories of all features:

$$\hat{C}(t) = \frac{\hat{C}(t-m) + \cdots + C(t) + \cdots + C(t+m)}{2m+1}. \quad (1)$$

Here $C(t)$ is the feature vector at frame t, $\hat{C}(t)$ is the filtered feature vector, and m is the order of the filter. We use a second order ARMA filter ($m = 2$), which we found consistently improves separation performance in low SNR conditions.

We use DNNs (multilayer perceptrons) as the discriminative learning machine, which has been shown to work well for speech separation [33], [34]. All DNNs use three hidden layers, each having 1024 rectified linear hidden units (ReLU) [7]. The standard backpropagation algorithm coupled with dropout regularization [14] (dropout rate 0.2) are used to train the networks. No unsupervised pretraining is used. We use the adaptive gradient descent [4] along with a momentum term as the optimization technique. A momentum rate of 0.5 is used for the first 5 epochs, after which the rate increases to and is kept as 0.9. The DNNs are trained to predict the desired outputs across all frequency bands, and the mean squared error (MSE) is used as the cost (loss) function. The dimensionality of the output layer depends on the target of interest, which is described in the next section. For targets in the range $[0, 1]$, we use sigmoid activation functions in the output layer; for the rest we use linear activation functions.

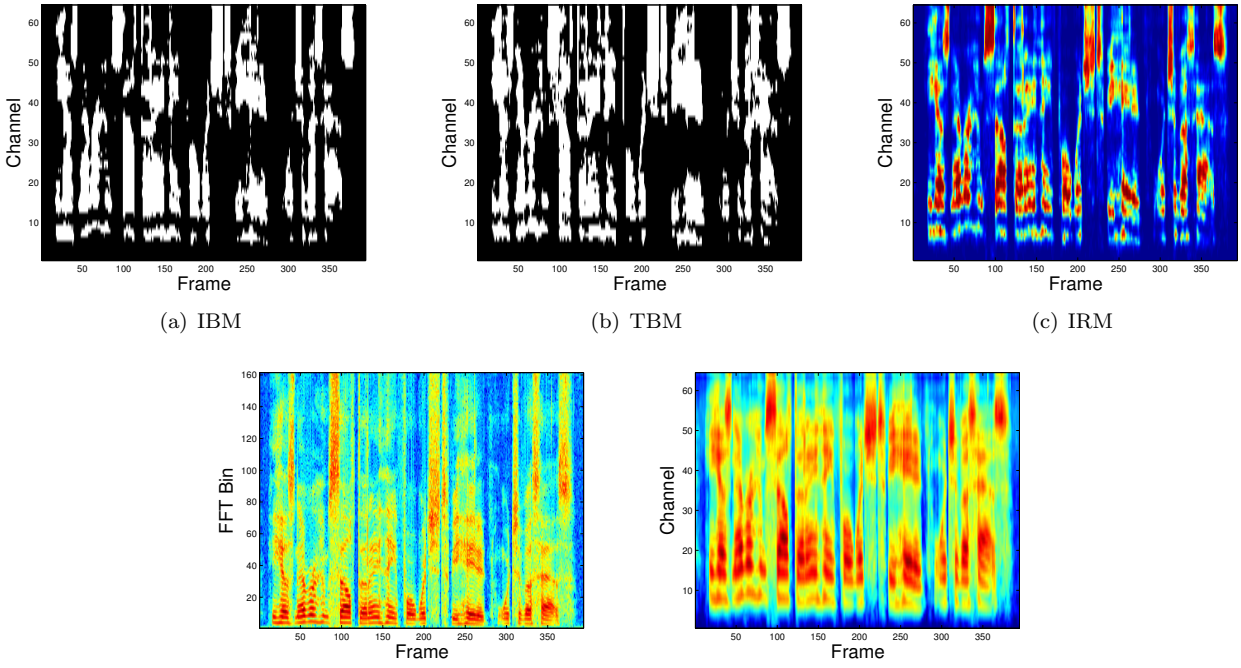
To further incorporate temporal context, we splice a 5-frame window of features as input to the DNNs. The output of the network is composed of the corresponding 5-frame window of targets. In other words, the DNNs predict the neighboring frames' targets together. The multiple estimates for each frame are then averaged to produce the final estimate. Doing so yields small but consistent improvements over predicting single-frame targets.

3 Training Targets

We introduce six training targets evaluated in this study below. We assume that the input signal is sampled at 16 kHz, and use a 20-ms analysis window with 10-ms overlap. An illustration of different training targets is shown in Figure 1.

3.1 Ideal Binary Mask (IBM)

The ideal binary mask is a main computational goal for computational auditory scene analysis (CASA) [30]. The IBM is a time-frequency (T-F) mask constructed from premixed signals. For each T-F unit, we set the corresponding mask value to 1 if the local SNR is greater than a local criterion (denoted as LC), otherwise it is set to 0. Quantitatively, the IBM is defined



(a) IBM (b) TBM (c) IRM

(d) FFT-MAG (e) GF-POW

Figure 1: Various training targets for a TIMIT utterance mixed with a factory noise at -5 dB SNR.

as:

$$IBM(t,f) = \begin{cases} 1, & \text{if } SNR(t,f) > LC \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $SNR(t,f)$ denotes the local SNR within the T-F unit at time t and frequency f. As mentioned earlier, it is well established that IBM processing of sound mixtures yields large speech intelligibility gains for both normal hearing and hearing impaired listeners. In addition, the effectiveness of IBM estimation has been exemplified by the recent success in improving human speech intelligibility [16], [11].

Following common practice, we use a 64-channel Gammatone filterbank to derive the IBM; hence the DNN has $64 \times 5 = 320$ sigmoidal output units with a 5-frame window of targets. Note that although we train on binary targets, in testing we use the posterior probabilities from the DNN, representing the probability of speech dominance, as a soft mask for resynthesis which is found to produce better quality. The choice of LC has a significant impact on speech intelligibility [17]; we set LC to be 5 dB smaller than the SNR of the mixture to preserve enough speech information. For example, if the mixture SNR is -5 dB, the corresponding LC is set to -10 dB.

3.2 Target Binary Mask (TBM)

Unlike the IBM, the TBM [17] is a binary mask that is obtained by comparing the target speech energy in each T-F unit with a reference speech-shaped noise (SSN). That is, the $SNR(t, f)$ term in Eq. (2) is calculated using the reference SSN, regardless of the actual interference. Although the TBM is obtained in a noise-independent way, subject tests have shown that the TBM achieves similar intelligibility improvements as the IBM [17]. The reason the TBM works is that it preserves the spectrotemporal modulation patterns essential to speech perception, i.e. where the target speech energy occurs in the time-frequency domain. Since the TBM can be interpreted as a cartoon of target speech patterns, it may be easier to learn the TBM than the IBM. We use the same frontend to generate the TBM, i.e. a 64-channel Gammatone filterbank, and the same LC values.

3.3 Ideal Ratio Mask (IRM)

The ideal ratio mask is defined as follows:

$$\begin{aligned} IRM(t,f) &= \left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta} \\ &= \left(\frac{SNR(t,f)}{SNR(t,f)+1}\right)^{\beta}, \end{aligned} \tag{3}$$

where $S^2(t, f)$ and $N^2(t, f)$ denote the speech and noise energy, respectively, in a particular T-F unit. β is a tunable parameter to scale the mask. Although technically different, one can see that the IRM is closely related to the frequency-domain Wiener filter assuming speech and noise are uncorrelated [26], [20]. We experimented with different β values and found $\beta = 0.5$ to be the best choice. Interestingly, with $\beta = 0.5$, Eq. (3) becomes similar to the square-root Wiener filter, which is the optimal estimator of the power spectrum [20].

Like the IBM and TBM, the IRM is also obtained by using a 64-channel Gammatone filterbank and is in the range of $[0, 1]$.

3.4 Gammatone Frequency Power Spectrum (GF-POW)

We also evaluate performance by directly predicting the 64-channel Gammatone frequency power spectrum (GF-POW) of clean speech. The Gammatone filterbank has a finer resolution in lower frequency regions compared to STFT. Since there is no direct inverse transformation for Gammatone filtering, we convert the estimated power spectrum to a mask, $\sqrt{S^2_{GF}(t,f)/Y^2_{GF}(t,f)}$, for resynthesis. Here, S^2_{GF} and Y^2_{GF} denote the speech and noise energy in the Gammatone frequency domain, respectively. Predicting GF-POW has been shown to be useful in our recent supervised dereverberation work [10]. Note that the construction of the IBM and the IRM also involves GF-POW.

3.5 Short-Time Fourier Transform Spectral Magnitude (FFT-MAG) and Mask (FFT-MASK)

If the goal is to recover clean speech, then predicting the STFT magnitude of clean speech seems to be a very natural choice. We use a 320-point FFT analysis; thus the spectral magnitude in each frame is a 161-D vector. The estimated magnitude combined with the mixture phase is passed through an inverse FFT to generate the estimated clean speech. Raw spectral magnitudes usually have a very large dynamic range, hence proper compression or normalization are needed to make them amenable to the backpropagation training. We will show in Section 5.1 that different normalizations can lead to significantly different results. We note that a very recent study [36] predicts the log compressed spectral magnitude under a supervised speech separation framework.

For a comparison with masking, we straightforwardly rewrite the clean spectral magnitude, $S_{FFT}(t, f)$, as

$$S_{FFT}(t, f) = \frac{S_{FFT}(t, f)}{Y_{FFT}(t, f)} \cdot Y_{FFT}(t, f), \tag{4}$$

where Y_{FFT} is the (uncompressed) noisy STFT spectral magnitude. Apart from directly predicting S_{FFT}, we also predict $S_{FFT}(t, f)/Y_{FFT}(t, f)$, which can be interpreted as a mask. The clean magnitude is reconstructed by multiplying the predicted mask with the noisy magnitude. We call this target FFT-MASK, which is an intermediate target to recover clean magnitude. Note that unlike the IRM, the FFT-MASK is not upper-bounded by 1, and therefore we use linear output activation functions in the DNNs. For better numerical stability in backpropagation training, we clip values greater than 10 in the FFT-MASK to 10, where 10 is an arbitrarily chosen value. The motivation of introducing FFT-MASK is to enable a direct comparison with FFT-MAG, as the perfect estimation of these two targets produces the same underlying objective – the clean speech magnitude.

4 Experimental Settings

We use 2000 randomly chosen utterances from the TIMIT [6] training set as our training utterances. The TIMIT core test set, which consists of 192 utterances from unseen speakers of both genders, is used as the test set. We use SSN and 4 other noises from the NOISEX dataset [28] as our training and test noises. These include a babble noise, a factory noise (called "factory1"), a destroyer engine room noise, and an operation room noise (called "oproom"). Except SSN, all other noises are non-stationary. All noises are around 4 minutes long. To create the training sets, we use random cuts from the first 2 minutes of each noise to mix with the training utterances at -5 and 0 dB SNR. The test mixtures are constructed by mixing random cuts from the last 2 minutes of each noise with the test utterances at -5, 0 and 5 dB SNR, where 5 dB is an unseen SNR condition. Dividing the noises into two halves ensures that

Table 1: Performance on factory1 when the clean magnitudes are normalized/compressed in different ways

Methods	**-5 dB**		**0 dB**		**5 dB**	
	STOI	PESQ	STOI	PESQ	STOI	PESQ
Log Compression	0.65	1.82	0.72	2.11	0.75	2.23
Percent Normalization	0.65	1.60	0.73	1.76	0.77	1.83
Log + Percent Norm.	0.66	1.73	0.75	2.06	0.79	2.25

the new noise segments are used during testing. Aside from the aforementioned noises, we also use an unseen factory noise (called "factory2") and a tank noise from NOISEX to evaluate generalization performance. Note that separation of these broadband noises at low SNRs is a very challenging task. For example, even for the stationary SSN, the human intelligibility score at -5 dB is only around 65% and 35% for normal hearing and hearing impaired listeners, respectively [11].

To put our results in perspective, we compare with model-based speech enhancement and NMF based separation. For speech enhancement, we compare with a recent system by Hendriks et al. [12], which uses an MMSE estimator of speech DFT coefficients assuming a generalized gamma distribution for speech magnitude [5]. For noise tracking, a state-of-the-art MMSE noise power spectral density estimator is used [12]. For a fair comparison with NMF, we use the supervised NMF method where the speech bases and noise bases are trained separately (for each type of noise) using the same training data used by the DNNs. We made an effort to yield the best performance of supervised NMF; we have tried different variants and found that a recent version using an active-set Newton algorithm (ASNA) [29] produces the best results. The final system, denoted as ASNA-NMF, models a sliding window of 5 frames of magnitude spectra and uses 160 speech bases and 80 noise bases. Using larger numbers of bases (e.g., 1000) does not seem to improve the performance significantly on our test set.

For evaluation metrics, we use the Short-Time Objective Intelligibility score (STOI) [27] to measure the objective intelligibility. STOI denotes a correlation of short-time temporal envelopes between clean and separated speech, and has been shown to be highly correlated to human speech intelligibility score. We also evaluate objective speech quality using the Perceptual Evaluation of Speech Quality (PESQ) score [25]. Like STOI, PESQ is obtained by comparing the separated speech with the corresponding clean speech. The STOI score ranges from 0 to 1, and PESQ score -0.5 to 4.5.

To supplement the above perceptually oriented metrics, we also give SNR results, which take into account the underlying signal energy of each T-F unit. We should point out that the traditional SNR metric comparing the separated speech with clean speech is not appropriate here. First, different targets aim to reconstruct different underlying signals. For example, the ground truth signal of IBM prediction differs from that of FFT-MASK prediction, therefore the use of the traditional SNR is problematic. Second, the traditional SNR does not take account of perceptual effects and it is well documented that SNR may not correlate with speech intelligibility. For example, $LC = 0$ dB maximizes the SNR gain of the IBM [19].

Table 2: Performance comparisons between various targets and systems on -5 dB mixtures. "MC-IRM" stands for multi-condition training (on all five noises) and uses IRM as the training target

Target/System	Factory1			Babble			SSN			Engine			Oproom		
	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR
Mixture	0.54	1.29	-5.00	0.55	1.42	-5.00	0.57	1.48	-5.00	0.57	1.41	-5.00	0.59	1.40	-5.00
IBM	0.66	1.49	6.63	0.63	1.50	3.98	0.72	1.45	8.71	0.78	1.53	13.24	0.77	1.81	12.24
TBM	0.65	1.33	5.19	0.62	1.32	3.08	0.72	1.45	8.71	0.77	1.52	6.16	0.76	1.60	6.38
IRM	0.67	1.75	8.27	0.63	1.64	4.39	0.73	**1.87**	10.81	**0.80**	**2.17**	15.66	**0.79**	**2.19**	15.33
FFT-MAG	0.66	1.73	5.45	0.62	1.50	3.80	0.72	1.76	5.18	0.76	2.02	6.09	0.74	2.01	5.84
FFT-MASK	**0.68**	1.77	7.59	**0.65**	**1.65**	5.52	**0.74**	**1.87**	7.58	0.78	2.16	9.73	0.77	2.15	9.89
GF-POW	0.67	**1.80**	8.23	0.62	1.63	5.98	0.72	1.85	8.62	0.76	2.06	9.83	0.74	2.14	9.31
MC-IRM	**0.69**	**1.80**	9.52	**0.64**	**1.65**	5.08	**0.74**	**1.88**	11.40	**0.78**	**2.12**	14.97	**0.77**	**2.16**	14.79
ASNA-NMF	0.60	1.55	5.62	0.57	1.53	4.21	0.64	1.61	5.69	0.70	1.84	7.04	0.68	1.81	7.08
SPEH	0.51	1.56	4.13	0.50	1.38	3.07	0.57	1.68	4.34	0.62	1.85	5.73	0.58	1.88	5.98

Table 3: Performance comparisons between various targets and systems on 0 dB mixtures

Target/System	Factory1			Babble			SSN			Engine			Oproom		
	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR
Mixture	0.65	1.62	0.00	0.66	1.73	0.00	0.69	1.75	0.00	0.68	1.66	0.00	0.70	1.78	0.00
IBM	0.78	1.85	12.17	0.76	1.89	8.91	0.82	1.75	14.48	0.85	1.79	17.13	0.83	2.11	16.13
TBM	0.77	1.67	10.43	0.75	1.65	8.20	0.82	1.75	14.48	0.84	1.75	9.76	0.82	1.83	10.08
IRM	0.78	2.17	14.03	0.76	2.05	10.54	**0.83**	2.26	16.39	**0.86**	2.48	19.23	**0.84**	**2.47**	18.62
FFT-MAG	0.75	2.06	6.09	0.72	1.89	4.92	0.78	2.09	5.71	0.80	2.28	6.36	0.78	2.23	6.21
FFT-MASK	**0.79**	**2.22**	11.07	**0.77**	**2.10**	9.44	**0.83**	1.87	10.59	0.85	**2.51**	12.14	0.83	**2.47**	12.43
GF-POW	0.76	2.19	9.68	0.73	2.05	8.09	0.80	2.23	9.64	0.81	2.38	10.40	0.80	2.42	10.19
MC-IRM	**0.79**	**2.20**	15.19	**0.77**	**2.07**	11.18	**0.83**	**2.26**	16.45	**0.85**	**2.43**	18.84	**0.83**	**2.45**	18.54
ASNA-NMF	0.72	1.93	9.15	0.71	1.91	7.64	0.76	1.97	8.86	0.80	2.19	9.81	0.77	2.15	10.14
SPEH	0.64	2.00	7.52	0.64	1.82	6.85	0.71	2.09	7.07	0.75	2.24	8.57	0.70	2.25	8.89

However, the choice of $LC = 0$ dB is clearly worse than negative LC values (e.g. -6 dB) for both human speech intelligibility [17] and automatic speech recognition performance [22]. In other words, lower output SNRs lead to higher speech intelligibility (see also [16], [11]). As our study focuses on different training targets, it makes sense to use the target-based SNR that compares the separated speech with the target signal resynthesized from the corresponding ideal target. That is, the output SNRs of IBM, TBM and IRM predictions are obtained using the signals resynthesized from the IBM, TBM and IRM, respectively, as the ground truth (see also [30]). For FFT-MAG, FFT-MASK, ASNA-NMF and Hendriks et al.'s system, the ground truth signal is resynthesized using the clean speech magnitude combined with the mixture phase, as the computational objective of these targets/methods is to obtain STFT clean speech magnitude and the separated speech is reconstructed using the mixture phase. Using the target-based SNR facilitates comparisons between these two groups of targets/methods.

5 Results

5.1 Comparison Between Targets

Before presenting comprehensive evaluations, we compare various compression/normalization techniques for predicting FFT-MAG. If one wants to predict the clean magnitude, proper normalizations or compressions are needed because the magnitudes typically have a very broad dynamic range, causing difficulty for gradient descent based training algorithms. We show

Table 4: Performance comparisons between various targets and systems on 5 dB mixtures

Target/System	**Factory1**			**Babble**			**SSN**			**Engine**			**Oproom**		
	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR	STOI	PESQ	SNR
Mixture	0.77	1.99	5.00	0.77	2.08	5.00	0.81	2.05	5.00	0.80	1.97	5.00	0.79	2.16	5.00
IBM	**0.86**	2.13	16.73	**0.86**	2.28	13.93	0.87	1.81	18.05	0.89	2.02	19.66	0.86	2.26	19.12
TBM	0.85	1.92	15.45	0.85	1.95	13.80	0.87	1.81	18.05	0.88	1.89	13.52	0.86	2.00	13.69
IRM	**0.86**	2.51	19.84	**0.86**	**2.47**	17.27	**0.88**	2.17	18.56	**0.90**	2.75	23.29	**0.88**	**2.77**	22.57
FFT-MAG	0.79	2.25	6.02	0.77	2.11	5.27	0.79	1.92	5.40	0.83	2.42	6.30	0.81	2.34	6.08
FFT-MASK	**0.86**	**2.59**	13.60	0.85	2.23	12.63	0.85	**2.23**	11.11	**0.90**	**2.80**	14.04	0.87	2.76	14.24
GF-POW	0.81	2.48	9.95	0.80	2.41	9.01	0.82	2.17	9.21	0.85	2.61	10.43	0.83	2.63	10.27
MC-IRM	**0.87**	**2.52**	20.65	**0.86**	**2.48**	17.87	**0.88**	2.33	19.98	**0.90**	**2.70**	22.92	**0.88**	**2.72**	21.98
ASNA-NMF	0.82	2.28	12.53	0.81	2.28	11.34	0.85	2.30	12.32	0.87	2.51	12.76	0.84	2.48	13.11
SPEH	0.77	2.39	10.84	0.76	2.25	10.66	0.82	**2.47**	10.25	0.85	2.59	11.55	0.80	2.62	11.51

that different ways of normalization impact performance significantly. In Table 1, we compare STOI and PESQ performance on the factory1 noise using different kinds of normalization/compression methods. We first use the log compression, which is perhaps the most widely used compression technique (e.g. [36]). Since log magnitude is not bounded, we use linear output units in the DNNs. Next, we use percent normalization, which linearly scales the data to the range of $[0, 1]$. This is done by first subtracting the minimum value, and then dividing by the difference between the maximum and minimum value. We use sigmoidal output units in this case. Finally, we normalize the magnitudes by first performing a log compression followed by percent normalization, and use sigmoidal output units. From Table 1 we can see that the performance of these normalization methods does not differ too much at -5 dB. However, at 0 and 5 dB, the traditional log compression performs significantly worse in terms of STOI (e.g., 4% worse at 5 dB) than the log compression followed by percent normalization. Using only percent normalization gives closer, but still worse, STOI results; but its PESQ results are the worst among the three. We believe log + percent normalization performs better because it preserves spectral details while simultaneously making the target bounded. Therefore, we use this normalization scheme when predicting spectral magnitude/energy based targets in the remaining experiments.

The comparisons between different targets in various test conditions are shown in Tables 2, 3, and 4 at different mixture SNRs, where best score is highlighted by boldface. In these tables, "SNR" denotes the target-based SNR mentioned in Section 4. We first discuss the results in the most challenging scenario, the -5 dB SNR case, as shown in Table 2. Generally, regardless of the target of choice, the supervised speech separation framework provides substantial improvements compared to unprocessed mixtures. For the two binary masking targets, the IBM and the TBM, large improvements are obtained in STOI and SNR. Although we use the posterior probabilities from the DNNs as soft masks for resynthesis, the PESQ improvements are still limited over unprocessed mixtures, except in the case of the operation room noise. This is consistent with the common point of view that binary masking tends to improve speech intelligibility but not speech quality. Compared to the IBM, using the TBM as the target results in similar STOI scores but significantly worse PESQ scores and SNRs.

For supervised techniques like the one used here, the IBM seems to be a better choice than the TBM, probably because the TBM is defined by completely ignoring the noise characteristics in the mixture.

Going from binary masking to ratio masking improves all objective metrics, as exemplified by the performance of the IRM. On factory1, babble and SSN, predicting the IRM achieves slightly better or equal STOI results than predicting the IBM. On the engine and operation room noises, predicting the IRM yields more than two percent STOI improvements. Predicting the IRM seems to be especially beneficial for improving objective speech quality. For example, the PESQ score improves by 0.65 and 0.76 in the engine noise compared to the IBM and unprocessed mixtures, respectively. On average, predicting the IRM provides a 2 dB SNR improvements over predicting the IBM.

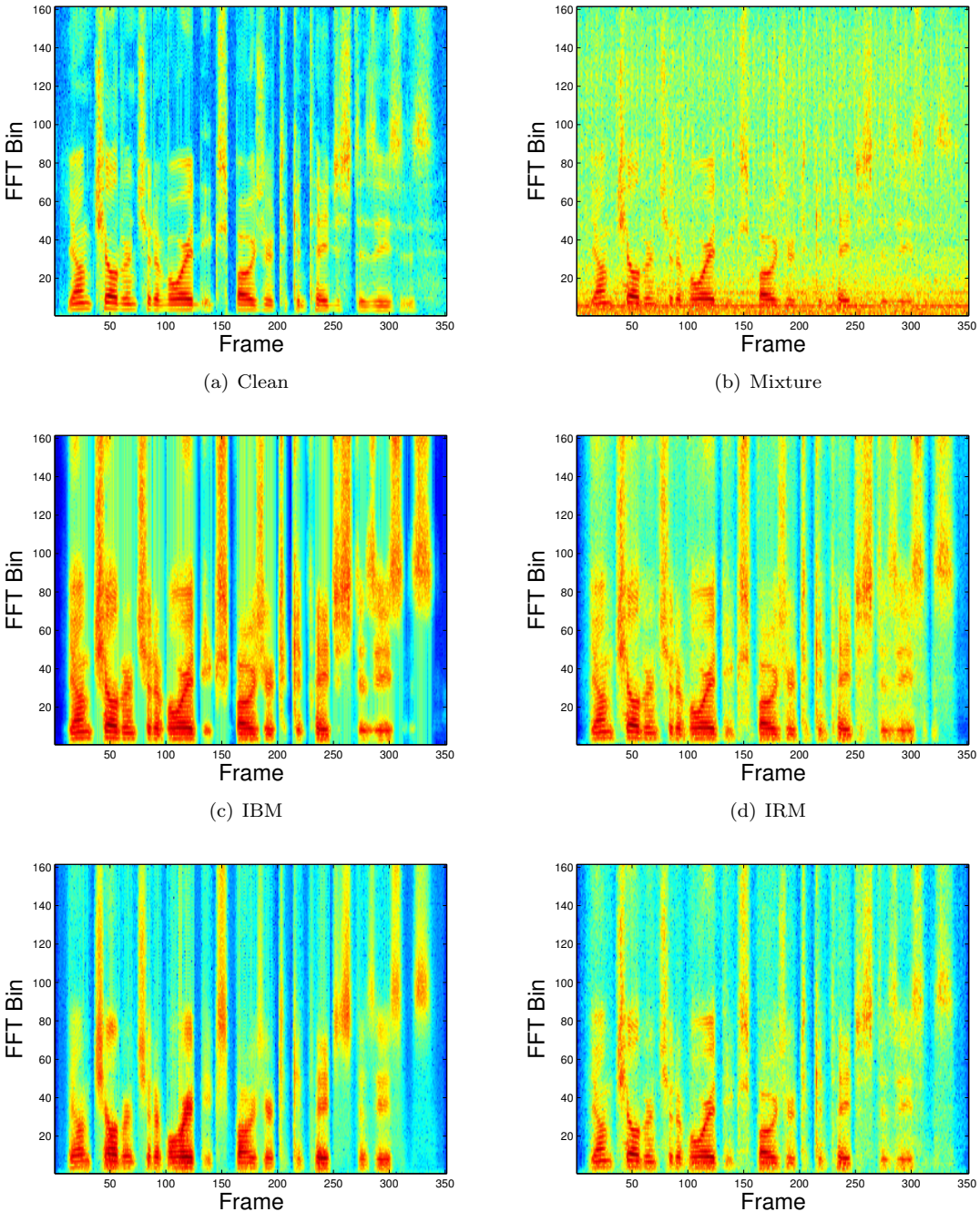
On the two challenging noises, factory1 and babble, and the relatively easier noise, SSN, predicting FFT-MAG achieves similar STOI and better PESQ results compared to predicting the IBM. However, predicting FFT-MAG achieves the worst performance on the other noises in terms of STOI. For example, on the operation room noise, the STOI is 3 percent worse than predicting the IBM. Similarly, FFT-MAG is consistently worse than the IRM, especially in the case of engine noise and operation room noise.

Interestingly, FFT-MASK produces comparable and sometimes even slightly better STOI and PESQ results than the IRM. Also, FFT-MASK produces significantly better SNR results than FFT-MAG on all noises. This contrast with FFT-MAG appears surprising at first, considering that the DNNs in both cases are essentially trained to estimate the same underlying target, the clean magnitude. We will provide some analysis in the next subsection as to why FFT-MASK performs better.

Predicting GF-POW, which is also a spectral envelope based target, has a similar performance trend as FFT-MAG. In general, it produces worse STOI results than either binary or ratio masking. Nevertheless, GF-POW seems to be consistently better than FFT-MAG.

The performance trend at 0 dB is similar to that at -5 dB, as shown in Table 3. That is to say, all targets improve objective metrics over unprocessed mixtures. Binary masking significantly improves objective intelligibility scores but the improvement in objective quality is minor. Predicting the IRM instead of the IBM significantly improves both objective quality and intelligibility metrics. FFT-MAG fails to compete with the other targets, whereas FFT-MASK is on par with the IRM. One noticeable difference at 0 dB is that the performance degradation of FFT-MAG becomes noticeably larger. For example, in the -5 dB factory1 noise condition, FFT-MAG produces the same STOI results as the IBM, whereas at 0 dB FFT-MAG is 3 percent points worse.

Separation at 5 dB is relatively easier, hence we can see in Table 4 that the STOI difference between various masking based targets becomes smaller. In contrast, FFT-MAG performs much worse than all the masking based targets. For example, the STOI and PESQ results obtained on SSN are even worse than those of unprocessed mixtures. In general, the IRM and



(a) Clean

(b) Mixture

(c) IBM

(d) IRM

(e) FFT-MAG

(f) FFT-MASK

Figure 2: STFT magnitudes of a separated speech using different training targets. The mixture here is a TIMIT male utterance mixed with the factory1 noise at 5 dB.

FFT-MASK perform comparably; with the former slightly better on average.

In Figure 2, we illustrate the STFT magnitudes of a separated speech utterance resynthesized using the estimated IBM, IRM, FFT-MAG and FFT-MASK. The mixture here is a TIMIT male utterance with factory1 at 5 dB. We can see that predicting the IBM preserves spectrotemporal modulation patterns of the clean speech, which are essential for speech intelligibility [23]. Also, the separated speech tends to have clearer onsets/offsets and sharper spectral transitions. Predicting FFT-MAG works reasonably well for low frequency regions where most of the speech energy resides. However, it misses a lot of details in the mid- to high-frequency regions, which are important for both intelligibility and quality. Visually speaking, the results are similar between IRM prediction and FFT-MASK prediction in the sense that they both preserve important modulation patterns as well as fine structures.

5.2 Why FFT-MAG Prediction Fails?

We are interested in why predicting FFT-MAG produces the worst results, and in particular, why there is a substantial performance gap between FFT-MAG and FFT-MASK. We start our analysis with two hypotheses. First, since FFT-MAG is the same across different noises and SNRs, the DNN has to learn a many-to-one mapping (recall that we train on -5 and 0 dB mixtures), which may be a more difficult task compared to learning a one-to-one mapping as in FFT-MASK. Second, masking may be inherently better. To verify these hypotheses, we designed two experiments. In the first experiment, we train a DNN to predict FFT-MAG only for 0 dB factory1 to reduce many-to-one mapping, and in the second experiment we train a DNN to predict $\log S / \log Y$ (LOGMASK), where S and Y denote the clean and noisy magnitude (with time and frequency index omitted), for -5 and 0 dB factory1. The STOI results are shown in Figure 3. From the figure we can see that learning a many-to-one mapping does not seem to be the cause as the performance in the matched SNR condition is only marginally better. Interestingly, the new mask $\log S / \log Y$ does improve performance over FFT-MAG, but is still significantly worse than FFT-MASK. This seems to indicate that, although masking is helpful, the use of log compression is likely the cause of the performance difference. We further analyze why log compression affects performance below.

Let η denote the ratio between the network's output and the desired output. Here $\eta \in [0, +\infty]$, and when $\eta > 1$ or $\eta < 1$, the neural network overestimates or underestimates the target, respectively. We assume that the network learns equally well for both targets, meaning that η is in the same range. In Figure 4, we plot the average η values obtained on the *training set* across frequency for both targets, and we can see that this is basically the case. Since the goal is to estimate the clean magnitude, we evaluate the estimation error in terms of the absolute deviation from the clean magnitude S. Recall that for FFT-MASK, we predict a mask S/Y, whereas for log magnitude we predict $\log S$. Therefore for FFT-MASK, the

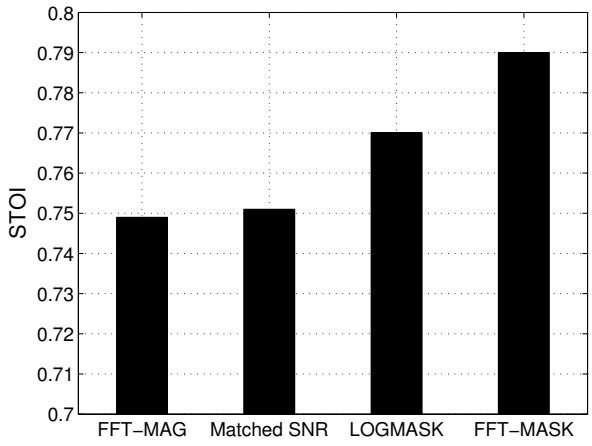


Figure 3: STOI results for 0 dB factory1. "Matched SNR" denotes training FFT-MAG also for 0 dB factory1. "LOGMASK" is a new mask ($\log S / \log Y$).

network output is $\eta S/Y$, and the estimation error E_{MASK} is:

$$\begin{aligned} E_{MASK} &= \left| S - \hat{S}_k \right| \\ &= \left| S - \eta \frac{S}{Y} Y \right| \\ &= |1 - \eta| S. \end{aligned} \tag{5}$$

For FFT-MAG, the network output is $\eta \log S$, and we have the estimation error E_{MAG} as:

$$\begin{aligned} E_{MAG} &= \left| S - \hat{S}_g \right| \\ &= |S - \exp(\eta \log S)| \\ &= |S - S^{\eta}| \\ &= \left| 1 - S^{\eta - 1} \right| S. \end{aligned} \tag{6}$$

In practice, $S \gg 1$ when speech is present, hence we assume $S > 1$. When $\eta > 1$ (overestimation), $E_{MAG} = (S^{\eta-1} - 1)S$, which is exponential with respect to η. In this case, E_{MAG} clearly grows much faster than E_{MASK}, which is linear with respect to η. For the case when $\eta < 1$, $E_{MASK} = (1 - \eta)S$ and $E_{MAG} = (1 - S^{\eta-1})S$. To see which error is greater, we plot in Figure 5(a) the contour of the ratio between E_{MAG} and E_{MASK} by varying S and η. We can see that E_{MASK} is consistently smaller than E_{MAG} except only when both S and η are very small (the region where the ratio is less than 1). This can also be seen in Figure 5(b), where we plot normalized error curves (E_{MASK}/S and E_{MAG}/S) when $S = 50$.

The analysis implies that, when using supervised techniques for estimating the clean magnitude, which is the underlying goal, one is likely more accurate by predicting the masking

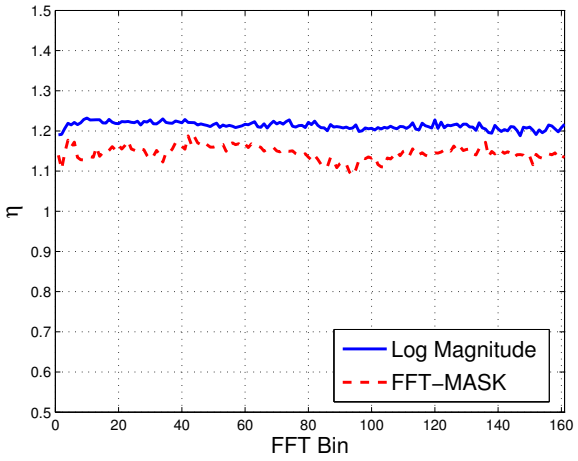
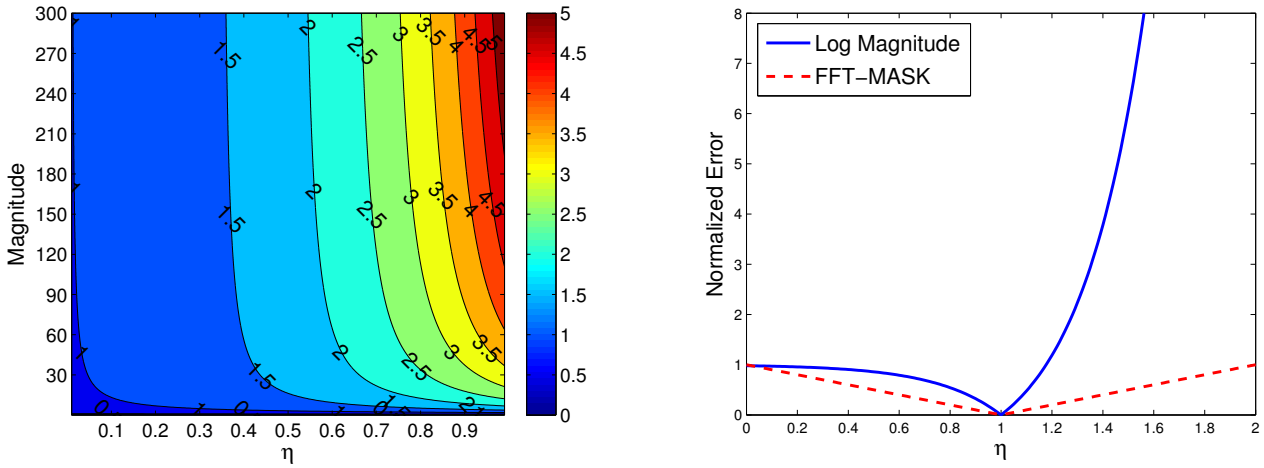


Figure 4: Average η values across frequency obtained on the training set.



(a) (b)

Figure 5: Comparison of E_{MAG} and E_{MASK}. Left: contour plot of the ratio E_{MAG}/E_{MASK}. Right: normalized error curves (E_{MASK}/S and E_{MAG}/S) when $S = 50$.

function FFT-MASK rather than the log magnitude. Roughly speaking, this is because the errors are magnified by the exponential expansion when converting to the magnitude domain before resynthesis. This analysis does not depend on the frequency scale, hence it also applies to the Gammatone frequency scale, partly explaining why predicting GF-POW is worse than predicting the IRM. In addition, we can show that similar analysis applies to other types of nonlinear compression, such as the cubic root compression $S^{1/3}$, and even to masks involving nonlinear compression, such as the LOGMASK mentioned above.

The log or root compression is often used for reducing the dynamic range. If one wants to predict the clean magnitude, such a compression is also needed in neural networks to avoid numerical issues so that gradients can flow well. Nevertheless, we argue that regardless of the

Table 5: Generalization performance on two unseen noises at -5 dB

Target	**Factory2**		**Tank**	
	STOI	PESQ	STOI	PESQ
Mixture	0.65	1.57	0.68	1.70
MC-IRM	**0.76**	**2.02**	**0.75**	**2.16**
ASNA-NMF	0.70	1.96	0.69	2.03
SPEH	0.65	1.95	0.68	2.04

Table 6: Generalization performance on two unseen noises at 0 dB

Target	**Factory2**		**Tank**	
	STOI	PESQ	STOI	PESQ
Mixture	0.76	1.93	0.77	2.06
MC-IRM	**0.85**	**2.44**	**0.83**	**2.52**
ASNA-NMF	0.80	2.33	0.79	2.38
SPEH	0.76	2.40	0.78	2.43

frequency scale, such a compression is better achieved via masking, which can be thought as a form of normalization, instead of a nonlinear compression.

5.3 Comparison with Other Systems

We now compare with a very recent supervised NMF system [29], i.e. ASNA-NMF, and a recent speech enhancement system [12], which we call SPEH. The results of these two systems are shown in the bottom of Tables 2, 3, and 4. ASNA-NMF is also a data driven method, trained on the same data as used by the DNNs. Its performance, however, is significantly worse than supervised speech separation. The performance difference in terms of STOI is particularly large compared to our DNN that predicts the IRM, e.g. 10 percent worse for the -5 dB engine noise. In challenging cases, e.g. -5 dB babble, ASNA-NMF improves upon unprocessed mixtures by only 2 percentage points in STOI. ASNA-NMF on average achieves significantly worse STOI but better PESQ results compared to binary masking (IBM and TBM). However, its PESQ results are consistently worse than ratio masking (IRM and FFT-MASK). Informal listening indicates that the output from ASNA-NMF has noticeable speech distortions and residual noise even at 5 dB. Speech enhancement, which does not rely on training, seems to have difficulty in challenging test conditions. SPEH fails to improve STOI for 4 out of 5 noises at -5 dB, and 3 out of 5 noises at 0 dB. Even at 5 dB where spectral patterns of speech are prominent, SPEH is outperformed by both DNN and ASNA-NMF. This is to be expected as the latter techniques are data-driven. With the same ground truth signal, FFT-MASK significantly outperforms ASNA-NMF and SPEH in terms of SNR. For example, at -5 dB, the average output SNR of FFT-MASK is 2.13 dB and 3.41 dB better than ASNA-NMF and SPEH, respectively.

In practice, supervised speech separation is often trained on multiple noises (i.e., multi-condition training) for good generalization (e.g. [34]). We train such a system on all 5 noises

Table 7: Generalization performance on two unseen noises at 5 dB

Target	**Factory2**		**Tank**	
	STOI	PESQ	STOI	PESQ
Mixture	0.84	2.29	0.85	2.41
MC-IRM	**0.90**	**2.80**	**0.89**	**2.88**
ASNA-NMF	0.86	2.64	0.86	2.70
SPEH	0.85	2.74	0.86	2.82

to predict the IRM. This system is called MC-IRM and its performance is also shown in Table 2 to 4. We can see that on average the performance does not degrade thanks to the representational power of DNNs. In fact, the performance of multi-condition training even improves over individually trained models sometimes (e.g., for -5 dB factory1), and the performance advantage over ASNA-NMF and SPEH remains.

We further compare the generalization performance on two unseen noises – a different factory noise (factory2) and a tank noise, both from NOISEX. We compare MC-IRM with ASNA-NMF and SPEH. The two data driven systems MC-IRM and ASNA-NMF are both trained on the previously used 5 noises at -5 and 0 dB. Note that the main purpose of this set of experiments is to compare relative performance. As we are training on only 5 noises, the presented results by no means represent the best obtainable ones for either MC-IRM or ASNA-NMF. It is expected that the performance will be significantly improved using more noises for training, at least for MC-IRM as indicated by the results in [34], [36]. The generalization results at -5, 0, and 5 dB are shown in Tables 5, 6, and 7, respectively. Note that at 5 dB, the SNR, speakers and noises are all new. We can see that MC-IRM again outperforms the other two systems on both noises across all SNR conditions, especially in terms of STOI. The STOI and PESQ improvements of MC-IRM over unprocessed mixtures are large, while the STOI improvements of the other two systems are limited or marginal. We should also point out that with more training data, separation in the test phase takes significantly more time for ASNA-NMF. In contrast, this does not affect supervised speech separation systems, where the additional computational burden occurs only in the training phase.

6 Concluding Remarks

Choosing a suitable training target is critical for supervised learning, as it is directly related to the underlying computational goal. In the context of speech separation, the speech resynthesized (either directly or indirectly) using any ideal target restores intelligibility and quality. In practice, however, since the targets have to be estimated, the choice should be made by considering how well it can be estimated and how the errors in estimation affect performance.

Traditionally, the IBM is used as the training target for supervised speech separation. Despite the simplicity of the IBM and the recent success of classification based speech separation it has inspired, it is unclear whether the IBM is the best target in terms of *estimation*. In this

study, we have systematically investigated the relative performance between various training targets, some new and others not, using both objective intelligibility and quality metrics. The compared targets can be categorized into binary masking based (IBM and TBM), ratio masking based (IRM and FFT-MASK), and spectral envelope based (FFT-MAG and GF-POW) targets. In general, we have found that binary masking produces worse objective quality results compared to ratio masking. We also found that binary masking leads to slightly worse objective intelligibility results than ratio masking. This is likely because predicting ratio targets is less sensitive to estimation errors than predicting binary targets. An unexpected finding of this study is that the direct prediction of spectral envelopes produces the worst results, as best illustrated by the substantial performance gap between FFT-MAG and FFT-MASK, where the two targets are essentially two alternative views of the same underlying goal, the clean speech magnitude. Aside from the analysis presented in Section 5.2, which points to the issue of nonlinear compression, we believe that masking has several advantages over spectral envelope estimation. Perhaps most importantly, masks make direct contact with the observed mixtures in the sense that they are used to modulate the mixtures in the time-frequency domain. In contrast, direct estimation of speech magnitude ignores the mixture which contains the true underlying signal. This can be problematic when there is a significant amount of erroneous estimation. Furthermore, ideal masks are inherently normalized and usually bounded, potentially making training easier and prediction more accurate compared to unbounded spectral envelope based targets.

We have also compared a supervised IRM estimation algorithm with recent algorithms in supervised NMF and statistical model-based speech enhancement. The comparisons across various test conditions clearly indicate significant performance advantage of our system.

To conclude, we hope efforts will be devoted to the design of new training targets in the future, which has the potential to further improve performance without adding significant computational burden. For example, predicting intermediate targets that encode more structure [35] or are easier to learn [8] has been shown to be useful.

Acknowledgements

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), an STTR subcontract from Kuzer, and the Ohio Supercomputer Center.

References

[1] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[2] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[3] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 257–270, 2007.

[4] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, pp. 2121–2159, 2011.

[5] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1741–1752, 2007.

[6] J. Garofolo, *DARPA TIMIT acoustic-phonetic continuous speech corpus*, National Inst. of Standards and Technology, 1993.

[7] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, vol. 15, 2011, pp. 315–323.

[8] C. Gulcehre and Y. Bengio, "Knowledge matters: Importance of prior information for optimization," in *International Conference on Learning Representations (ICLR)*, 2013.

[9] K. Han and D. Wang, "A classification based approach to speech segregation," *Journal of the Acoustical Society of America*, vol. 132, pp. 3475–3483, 2012.

[10] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," *in Proc. ICASSP, to appear*, 2014.

[11] E. Healy, S. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America*, pp. 3029–3038, 2013.

[12] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. ICASSP*, 2010, pp. 4266–4269.

[13] 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*, pp. 45–66, 2010.

[14] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[15] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 625–638, 2009.

[16] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, pp. 1486–1494, 2009.

[17] U. Kjems, J. Boldt, M. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *Journal of the Acoustical Society of America*, vol. 126, pp. 1415–1426, 2009.

[18] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[19] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, pp. 230–239, 2009.

[20] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[22] ——, "The role of binary mask patterns in automatic speech recognition in background noise," *Journal of the Acoustical Society of America*, pp. 3083–3093, 2013.

[23] R. Plomp, *The intelligent ear: On the nature of sound perception*. Lawrence Erlbaum Associates Mahwah, NJ, 2002.

[24] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1766–1776, 2007.

[25] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[26] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[27] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 2125–2136, 2011.

[28] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[29] T. Virtanen, J. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 2277–2289, 2013.

[30] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Divenyi P., Ed. Kluwer Academic, Norwell MA., 2005, pp. 181–197.

[31] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 125, pp. 2336–2347, 2009.

[32] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 270–279, 2013.

[33] Y. Wang and D. Wang, "Cocktail party processing via structured prediction," in *Proc. NIPS*, 2012, pp. 224–232.

[34] ——, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1381–1390, 2013.

[35] ——, "A structure-preserving training target for supervised speech separation," *in Proc. ICASSP, to appear*, 2014.

[36] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, pp. 66–68, 2014.