A Direct Masking Approach to Robust ASR

William Hartmann, *Member, IEEE*, Arun Narayanan, *Student Member, IEEE*,
Eric Fosler-Lussier, *Senior Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Recently, much work has been devoted to the computation of binary masks for speech segregation. Conventional wisdom in the field of ASR holds that these binary masks cannot be used directly; the missing energy significantly affects the calculation of the cepstral features commonly used in ASR. We show that this commonly held belief may be a misconception; we demonstrate the effectiveness of directly using the masked data on both a small and large vocabulary dataset. In fact, this approach, which we term the direct masking approach, performs comparably to two previously proposed missing feature techniques. We also investigate the reasons why other researchers may have not come to this conclusion; variance normalization of the features is a significant factor in performance. This work suggests a much better baseline than unenhanced speech for future work in missing feature ASR.**

***Index Terms*—Direct masking, ideal binary mask, robust automatic speech recognition.**

I. Introduction

Baseline systems provide a benchmark for judging advances in a field. New methods and techniques are considered effective when their performance exceeds the baseline. As a field progresses, baselines slowly improve to reflect the advances that have been made. However, there comes a time to reevaluate the baseline altogether. For missing feature techniques designed to provide robustness in automatic speech recognition (ASR), systems typically have been compared against unmodified, non-robust features. In this work, we show that a stronger baseline is available that is comparable to or outperforms standard missing feature techniques. The paper also explores explanations for why this stronger baseline has been overlooked in the literature.

ASR has long been known to suffer from the presence of background noise [1]. Many techniques have been developed that attempt to address this issue. Model-based techniques incorporate noise models into recognition [2]; these techniques typically require some statistical knowledge about the noise source and may require modifications to the recognizer. Noise-robust features, on the other hand, attempt to maintain invariance of the calculated features regardless of the noise condition [3]. Speech enhancement instead attempts to remove the noise from the signal prior to feature calculation. These methods typically do not require modifications to the standard recognition system.

Traditional speech enhancement methods, such as spectral subtraction [4], attempt to modify frame-level noisy speech spectra to make them closer to those of clean speech. In this work, we focus on the Computational Auditory Scene Analysis (CASA) based approach to speech enhancement. CASA, as we consider it in this paper, refers to sound segregation based on the perceptual process of auditory scene analysis proposed by Bregman [5]. It typically operates on a time-frequency (T-F) representation of the input, and produces an output that can be viewed as a binary T-F mask.

One proposed goal of CASA is the ideal binary mask (IBM) [6]. Conceptually, the IBM is very simple. A signal is first transformed into a spectrotemporal representation; the spectrogram and cochleagram are two common examples. Each pixel of this two-dimensional image of the signal, or T-F unit, represents the amount of energy at a particular frequency and time. The IBM is the binary segregation of these pixels into two groups; one containing energy mostly from a target source and one containing energy mostly from the interference. Formally, we can define the IBM as

$$M(\omega,t)=\begin{cases}1 & \frac{|S(\omega,t)|^2}{|N(\omega,t)|^2}>\theta\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

where ω is a frequency band and t represents a particular time frame. $S(\omega,t)$ and $N(\omega,t)$ represent the amount of energy at a T-F unit for the clean speech and interfering noise respectively. Fig. 1 provides an example. The threshold θ is typically set to 1, corresponding to a local SNR of 0 dB. We note that other studies have found improved performance using alternative thresholds [7], [8].

The IBM segregates the signal where a value of unity indicates that the corresponding T-F unit is grouped into the segregated target, and a value of zero indicates that the unit is considered part of the interference and hence removed [9]–[11]. We call T-F units with value 1 *unmasked*, and those with value 0 *masked*. Approaching the problem in this manner reduces speech enhancement to a binary classification task [6]. Previous studies have shown that processing noisy speech using an IBM can significantly improve speech intelligibility for humans (e.g. [12]).

Manuscript received May 17, 2012; revised September 14, 2012, February 13, 2013, and April 23, 2013; accepted May 03, 2013. Date of publication May 17, 2013; date of current version July 22, 2013. The work of W. Hartmann and E. Fosler-Lussier was supported in part by the National Science Foundation under Carrier Grant ISI-0643901. The work of A. Narayanan and D. Wang was supported in part by the Air Force Office of Scientific Research under Grant FA9550-12-1-0130. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Brian Kingsbury.

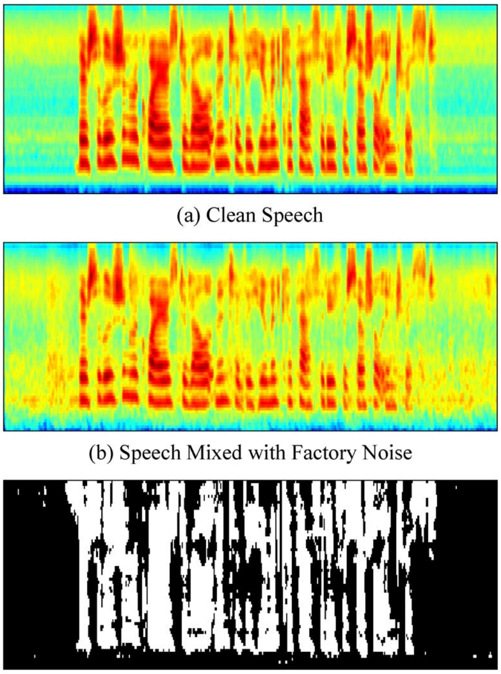
W. Hartmann, A. Narayanan, and E. Fosler-Lussier are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: hartmann.59@osu.edu).

D. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2263802

1558-7916/$31.00 © 2013 IEEE



(a) Clean Speech

(b) Speech Mixed with Factory Noise

(c) Ideal Binary Mask

Fig. 1. Ideal Binary Mask example. The top panel is a cochleagram representation of clean speech. The middle panel is that same speech mixed with factory noise at an SNR of 10 dB. The bottom panel is the IBM generated from the mixed speech. (a) Clean Speech; (b) Speech Mixed with Factory Noise; (c) Ideal Binary Mask.

For ASR, conventional wisdom in the field holds that the IBM cannot be used directly as the missing energy in the masked regions significantly affects the calculation of cepstral features. Directly using a mask refers to multiplying a spectrotemporal representation of the signal by the mask and then resynthesizing the waveform or calculating features from the masked signal. Based on this conventional wisdom, many techniques have been proposed to compensate for this missing energy when incorporating the IBM or related binary masks in ASR [13]–[15]. In this work, we will demonstrate that this conventional wisdom may, in fact, be a misconception. We extend our previous work in [16], which showed the IBM could be used directly in ASR for large-vocabulary tasks, in several ways. We present experiments on a small vocabulary dataset to show the effect is not due entirely to the strength of the language model. Our results show that directly using the IBM, which we term the *direct masking* approach[1] as we perform no reconstruction before resynthesis, not only works, but outperforms two main methods originally proposed to overcome the supposed inadequacies of the direct masking approach. In addition, we explain the likely cause for the original misconception—the lack of variance normalization in the features used. Results are also shown using estimated masks to demonstrate the conclusions hold in more realistic environments.

The rest of the paper is organized as follows. Section II presents background on the research incorporating the IBM in ASR. We describe in more detail two common approaches to utilizing the IBM in ASR and the direct masking approach in Section III. Our experiments on both small and large vocabulary tasks in Section IV show that using the IBM directly outperforms these two approaches in many cases. An analysis of why our results may have differed from previous work and our final conclusions are presented in Sections V and VI respectively.

II. Background

Once speech has been segregated using a binary mask, the question of how to perform ASR on the segregated speech still remains. Probably the first study to address this question is by Cooke *et al.* [18] who noted that standard ASR techniques had to be adapted to deal with occluded speech in masked T-F units. By treating masked speech as missing during classification, they adapted missing feature techniques in machine learning (e.g. [19]) to perform HMM based ASR, where the key idea is to marginalize over missing or unreliable features in probability calculation (see also [20]). Early studies demonstrated the effectiveness of marginalizing missing data or features by removing feature components either in the spectral domain [18], [20] or in the cepstral domain [18], which is far superior to recognizing noisy speech without processing. Obviously, to apply missing feature recognition in practice requires a missing feature detector that provides binary labels on feature components such as T-F units, which is considered the task of speech segregation. Since speech segregation algorithms operate in the time-frequency (spectral) domain, not in the cepstral domain, missing feature recognition has focused on coupling with HMM recognizers using spectral features [13].

It is well known that, for ASR, spectral or T-F features are not nearly as effective as cepstral features [21]. The success of marginalization has been mainly demonstrated on small-vocabulary tasks, such as digits or phones, and its scalability to larger vocabularies is questionable [22]. Treating marginalization as classifier compensation, Raj *et al.* [14] proposes feature compensation techniques by reconstructing missing features in the spectral domain based on a prior model of speech. With missing features reconstructed, a whole spectral vector can then be converted to the cepstral domain where conventional ASR is performed. A later study by Srinivasan *et al.* [23] additionally converts spectral uncertainty in mask estimation to the cepstral domain for improved ASR. Others advocate ratio masks, akin to Wiener filtering in speech enhancement, in place of binary masks in order to couple with cepstral features [22], [24]. Soft-masking [25], [26] takes a similar direction; the mask is altered by either the estimated energy or the uncertainty in the estimate. Ratio masks, soft masks, and reconstruction are all similar approaches in the sense that they use a real-value mask instead of a binary mask in the final enhancement.

Reconstruction of unreliable features from the reliable portion of the signal is an inference that is inherently error-prone. It hence seems logical to ask the question: What if no reconstruction is attempted? We have found only one study that explored the results of this simple direct masking approach. Cooke *et al.* [17] found that not modifying the masked regions performed significantly worse than any other approach when using spectral features. In addition, there are reasons to doubt the utility of

[1] Previous work has referred to this approach as zero-imputation [17]. We believe the term direct masking better distinguishes the approach from other techniques as no explicit compensation for the missing energy is made.

the direct masking approach. First, conventional wisdom would suggest that something must be done to holes (zeros) in a spectrogram or cochleagram created by binary masking [18]. This conventional wisdom is well founded when recognition is performed in the spectral domain as marginalization is a theoretically optimal technique. Second, it is not unreasonable to think that reconstruction, despite its approximate nature, should beat no reconstruction at all. This reasoning is encouraged by generally good results obtained from reconstruction research in comparison to recognition of noisy speech or some enhanced version via e.g. spectral subtraction [14], [23], [27]. Subsequent studies likely either ignored this approach based on the results in [17] or obtained poor results and declined to report them. Nonetheless, one would think that the condition ought to be included as a baseline in comparisons. We will make those comparisons in this study and present a likely reason for its absence in previous work.

III. Common Approaches

In the previous section, we gave a brief overview of the research concerning the incorporation of binary masks in ASR. We now present a more detailed description of marginalization and reconstruction approaches. Both techniques are examples of missing feature ASR. We separate the two techniques by how they handle missing features; for a more detailed review, see [7]. We also discuss the simplest method of incorporating the binary mask, directly using the masked signal.

A. Marginalization-Based ASR

Originally proposed by Cooke *et al.*, marginalization [13] was the first approach to address the issue of incorporating binary masks in ASR. While several variations were described in [13], we will focus here on the best performing method—bounded marginalization. Features are partitioned into reliable and unreliable ones based on a binary mask. Masked T-F units correspond to unreliable and unmasked units to reliable features. The marginalization-based speech recognizer is a modified HMM-GMM based speech recognizer that treats these masked and unmasked units in separate ways.

In a typical HMM based recognizer, every state is modeled by a GMM. The likelihood of a feature vector X given a particular state Q_i can be obtained by evaluating $p(X|Q_i)$. By separating the feature vector into reliable and unreliable components, the evaluation becomes

$$p(X|Q_i) = \int p(X_r, X_u|Q_i)dX_u \quad (2)$$

where we integrate over (i.e. marginalize) the possible values of X_u. As we are using a GMM with diagonal covariance for modeling, this becomes

$$p(X|Q_i) = \sum_{c=1}^{M} p(c|Q_i)p(X_r|c, Q_i) \int p(X_u|c, Q_i)dX_u \quad (3)$$

where c is a particular Gaussian and M is the number of Gaussians in the GMM. The assumption X_r and X_u are independent given c is implicitly made. While this is not true in practice, it follows from the use of a diagonal covariance matrix in the Gaussians.

Just as we partitioned the feature vector into reliable and unreliable portions, we can partition the means and variances of each Gaussian. We can then evaluate $p(X_r|c, Q_i)$ by evaluating the Gaussian only over the reliable dimensions. If we do not assume anything about the unreliable data, then the integral evaluates to one. However, we can at least determine bounds of the true feature based on the unreliable vector. Assuming X represents speech energy and we ignore phase interactions, then the true speech cannot have negative energy or more energy than in X_u. Note that while this assumption can sometimes be violated due to phase interactions, this effect is commonly ignored in missing data literature. The integral can then be evaluated using these bounds for a more accurate result.

Assuming that a given binary mask is accurate, the marginalization-based recognizer utilizes the available information from all the T-F units. Reliable units are treated in the standard way and unreliable features provide bounds on marginalization. On small vocabulary tasks such as TIDigits [28], the marginalization approach performs remarkably well. However, performance on larger vocabulary systems degrades significantly [14], [22]. A likely cause is the use of spectral features instead of the cepstral features which are known to perform better in ASR [21]. Methods that allowed for the calculation of cepstral features were needed to further increase performance, at least for larger vocabularies.

B. Reconstruction-Based ASR

One method that allows for the calculation of cepstral features is the estimation or reconstruction of missing T-F units. If the missing T-F units can be reconstructed, then the zeros or holes in the spectral representation no longer present a problem for cepstral feature calculation. The first comprehensive study of feature reconstruction was presented by Raj *et al.* [14].

It was clearly shown that this method only provided improvements over marginalization when using cepstral features. If instead the recognition was performed in the spectral domain, the reconstructed features performed worse. The results also held over larger vocabulary tasks. One benefit of this technique is that it does not require any modification to a standard recognizer.

Many specific techniques for performing the reconstruction have been explored [14], [15], [29]. We will present a technique that has been previously shown to improve results over a baseline system [23]. A comparison between this method and directly using the IBM will allow us to determine if reconstruction is always the preferred approach. We note that while the reconstruction approach in [23] is similar to the cluster-based approach proposed on [14], it does not make use of boundary information when calculating the posterior probability for each Gaussian component.

As with marginalization, a binary mask is used to partition the noisy speech vector Y into a reliable set Y_r and an unreliable set Y_u where $Y = Y_r \cup Y_u$ and $Y_r \cap Y_u = \varnothing$. Given Y, we want to estimate the true spectral vector $\hat{X}$ for the clean speech.

Assume $X_r = Y_r$. In order to estimate X_u, a speech prior is used [14]. The speech prior, consisting of spectral features instead of the cepstral features eventually used for recognition, is modeled by a GMM. Note that while this approach can use full covariance matrices, we use diagonal covariance matrices in our experiments. Just as we used the binary mask to partition the spectral vector, we can also use it to partition the mean and covariance of each mixture.

$$\mu_c = \begin{bmatrix} \mu_{r,c} \\ \mu_{u,c} \end{bmatrix} \quad \Sigma_c = \begin{bmatrix} \Sigma_{rr,c} & \Sigma_{ru,c} \\ \Sigma_{ur,c} & \Sigma_{uu,c} \end{bmatrix} \tag{4}$$

Ideally we would select the Gaussian that generated the spectral vector for estimation. Since we cannot identify the specific Gaussian, the estimate is the weighted sum of the estimates from each Gaussian.

$$\hat{X}_u = \sum_{c=1}^{M} p(c|X_r)\hat{X}_{u,c} \tag{5}$$

where M is the number of Gaussians and $\hat{X}_{u,c}$ is the expected value of X given the cth Gaussian. To estimate $p(c|X_r)$, the marginal distribution $p(X_r|c) = N(X_r; \mu_{r,c}, \Sigma_{rr,c})$ is used [23]. Finally, we compute the expected value of X_u given the cth Gaussian by

$$\hat{X}_{u,c} = \mu_{u,c} + \Sigma_{ur,c}\Sigma_{rr,c}^{-1}(X_r - \mu_{r,c}). \tag{6}$$

The unreliable portion of the spectral vector is then replaced by the estimate $\hat{X}_u$ and cepstral features are computed from the reconstructed spectrogram. If the estimate $\hat{X}_u$ exceeds the original value Y_u, then the original value is kept instead. Again, while this formulation can make use of a prior using full covariance matrices [14], our experiments, as in our previous work [16], use diagonal covariance matrices. The use of diagonal covariance matrices limits the amount of cross-channel interaction that can be leveraged during reconstruction.

Given that this approach allows for the calculation of cepstral features, performance is expected to scale to any size vocabulary. As methods for improving reconstruction further develop (e.g. [27], [30]), ASR results should also improve. While all of these techniques allow for the incorporation of a binary mask in ASR and show strong improvements over the baseline of recognizing noisy speech directly, they began with the implicit assumption that a binary mask cannot be used directly in ASR.

C. Direct Masking Approach

Both of the previously mentioned approaches start from the same point. Given a binary mask, estimated or ideal, they mask the signal in the spectral domain. From this point, marginalization marginalizes over the masked regions and reconstruction estimates the original signal in the masked regions. However, there is a simple third approach that treats the masked regions as having zero energy. The direct masking approach begins from the same point as the other two methods, but makes no attempt to compensate for the masked regions. Other work has referred to it as zero-imputation [17], but we feel the term direct masking better captures the distinction between this approach and the previous two as no actual imputation or marginalization takes place.

Given a computed binary mask in a spectral domain, we multiply the mask by the representation of the signal in that same domain. From the masked spectral representation of the speech, we can either calculate features or resynthesize a waveform signal. In our study, we resynthesize the waveform signal prior to calculating features, but directly calculating features from the masked spectral representation does not significantly change the results. Resynthesizing the waveform also allows mask estimation and feature calculation to be done in different signal representations.

The difference between this method and the previous two approaches is that no attempt is made to compensate the signal, features, or recognition system for the artificial zeros introduced in the spectral domain. One possible issue is that many features require a log operation and the log of zero is undefined. However, the artificial zeros are typically not a problem because a small amount of dither or noise is added to the signal before the log operation is performed in standard MFCC and PLP calculation software. Alternatively, a value other than 0 can be used for masked units in the binary mask. Small values produce similar results to using a value of 0 when using ideal binary masks, however, as will be shown in the next section, nonzero values for the masked units are essential for improved performance when using estimated masks.

Outside of Cooke *et al.* [17], we have been unable to find a study using this simple approach. The likely cause is that results were very poor when using this approach in [17]. Since that study, the conventional wisdom has been that this direct masking approach does not work. Given this assumption, much investigation has been made into alternative methods of utilizing the binary mask in ASR. We examine the validity of this assumption in the next section.

IV. Re-Evaluation Experiments

Our experiments parallel the research described in the previous section. We will compare the results of bounded marginalization and spectral reconstruction to the direct masking approach on both a small and large vocabulary dataset. The small vocabulary dataset, TIDigits [28], was used in Cooke *et al.* [13]. Since the strength of the spectral reconstruction technique over marginalization was seen on larger datasets, we will focus on spectral reconstruction for the large vocabulary dataset, Aurora4 [31]. Our recognition systems are similar to the ML-trained systems used in [13], [14], [23]. We acknowledge that our baseline results are not state of the art, but our goal was to have a fair comparison to the previously discussed studies.

A. TIDigits

The TIDigits corpus [28] consists of connected digit utterances. It has been widely used for speaker independent ASR studies [13], [32], [33]. As in previous studies, we use the male and female subsets of the corpus. The training set consists of 8623 utterances spoken by 55 male and 57 female speakers. The test set consists of 8700 utterances by a different set of 56 male and 57 female speakers, making the task speaker and gender independent. Note that, unlike the original study on missing feature recognition by Cooke *et al.* [13], we use the full test set to evaluate different ASR strategies. The direct masking approach is compared with marginalization and reconstruction based missing data approaches.

Before presenting our results, we first describe the features and models used. Marginalization based recognition is typically performed in the T-F domain. Since features in the T-F domain can be defined in multiple ways, we choose 5 more commonly used feature representations to evaluate marginalization. They are as follows:

- Cochleagram: Cochleagram is a popular feature representation in CASA that has been widely used for IBM estimation and other purposes [34]. To generate cochleagram based features, the signal is first passed through a 64 channel gammatone filterbank to perform T-F decomposition [34] (see Ch. 1). The channel center frequencies are uniformly spaced from 50 Hz to 8000 Hz in the ERB-rate scale. The output at each channel is then windowed using a 20 msec rectangular window with a 10 msec overlap (this corresponds to a frame rate of 100 Hz). The energy within each window is finally compressed using a log operation to obtain the cochleagram based feature at each T-F unit. In order to bound the feature values from below by 0, the energy values are incremented by 1 before compression.
- Rate64: The Rate64 features are obtained in a similar fashion. After decomposing the signal using a 64-channel gammatone filterbank, the instantaneous Hilbert envelope is extracted at the output of each channel. The envelope is then smoothed using a first-order filter with a time constant of 8 msec and downsampled to 100 Hz to obtain the features. Rate64 features are used in [13]. We use the CASA Toolkit [35] to extract this feature.
- Cubic compressed Rate64 (CRate64): This feature is similar to Rate64. After the initial T-F decomposition, the Hilbert envelope at each channel is directly downsampled to 100 Hz without smoothing, followed by a cubic root compression operation. Cubic compressed ratemap features are used in [33]. Smoothed versions have also been used in other studies [25], [32].
- Cubic compressed Rate64 with delta (CRate64_D): Studies in marginalization-based ASR have shown that adding delta components (temporal derivatives) can be useful in improving ASR performance [25], [36]. Therefore, as a fourth feature, we augment CRate64 features with their temporal derivatives to obtain CRate64_D features. We chose CRate64 because it produced the best performance on a smaller development set of 240 utterances.
- Spectrogram: All the above feature representations use a non-linear frequency axis. As a fifth feature, we use the spectrogram representation that has a linear frequency axis. Spectrogram features are obtained by first transforming the time-domain signal to the spectral domain using the FFT. The frame rate is set to 10 msec and the window size to 20 msec. A Hamming window is used, as is commonly done. The energy (squared amplitude) within each T-F unit is finally compressed using the log operator, as in the case of cochleagram features, to obtain a 160 dimensional feature representation at each time frame. Spectrogram based features are used in [22]. We did not add delta components since the performance with delta components was found to be comparable to those without them, when tested on the smaller development set.

For the direct masking and reconstruction based approaches, mean and variance normalized perceptual linear predictive (PLP) cepstral coefficients are extracted from the segregated target signal to perform recognition. A 39-dimensional feature representation that consists of 13 static coefficients along with its delta and acceleration coefficients are used. Segregation is performed either in the linear frequency domain using spectrogram features, or the non-linear frequency domain using cochleagram features. When using the spectrogram representation, the target is resynthesized from the mixture using the inverse DFT and the overlap-add method. Before applying the inverse DFT, the unreliable (masked) values of the spectrogram, as defined by the IBM, are set to 0 in the direct masking approach. In the reconstruction based approach, they are estimated using the method described in the previous section. A 1024-component, GMM-based speech prior model is trained using the training set of the TIDigits corpus for this purpose. When using the cochleagram representation, the target signal is resynthesized from the mixture using the method described in [34] (see Section 1.3.6), which is based on an approach introduced by Weintraub [37]. For the direct masking approach, the IBM is used directly to segregate the target speech. For the reconstruction based approach, masked T-F units of a cochleagram are first reconstructed. The reconstructed feature value in each T-F unit is then used to determine the percentage of target speech energy with respect to the mixture energy within the unit. Together with the 1s in the IBM, this defines a ratio mask for the mixture signal which is then used to resynthesize the target [34], [37]. Unlike the spectrogram, resynthesis in the cochleagram domain is defined in terms of the original signal and a ratio mask.

In all three approaches, the IBM defined using a local SNR criterion of 0 dB is used to identify the masked and unmasked regions in the T-F representation of a noisy utterance [6].[2] The 0 dB criterion is commonly used in CASA to define binary masks. The IBM is defined for each feature separately, by comparing the premixed target and noise energy at each T-F unit. Even though the masks look strikingly similar for all features, they do have some differences. The delta mask for CRate64_D is defined in the same way as in [25]. A delta feature is considered reliable or unmasked if all the static features used to calculate it are reliable, in accordance with the IBM. The direct masking and reconstruction based strategies use the IBMs corresponding to the cochleagram and the spectrogram features.

The ASR module consists of 11 word-level HMMs, one for each digit (1–9, ‘oh’ and ‘zero’), a silence model, and a short-pause model. Each word-level HMM consists of 8 emitting states, with the observation probability modeled as a mixture of 10 diagonal Gaussian components [13], [23]. The short-pause model has only 1 state, tied to the middle state of

[2] Note that Cooke *et al.* [13] use a binary mask called the *a priori* mask based on whether the mixture energy is within 3 dB of the target energy, corresponding to a local criterion of 7.7 dB instead of 0 dB. We experimented with both the *a priori* mask and the IBM using a smaller development set of 240 utterances and found that the latter works better, and hence is used for marginalization-based ASR.

TABLE I
WORD ACCURACIES OBTAINED USING THE CLEAN TEST SET OF THE TIDIGITS CORPUS FOR VARIOUS FEATURES

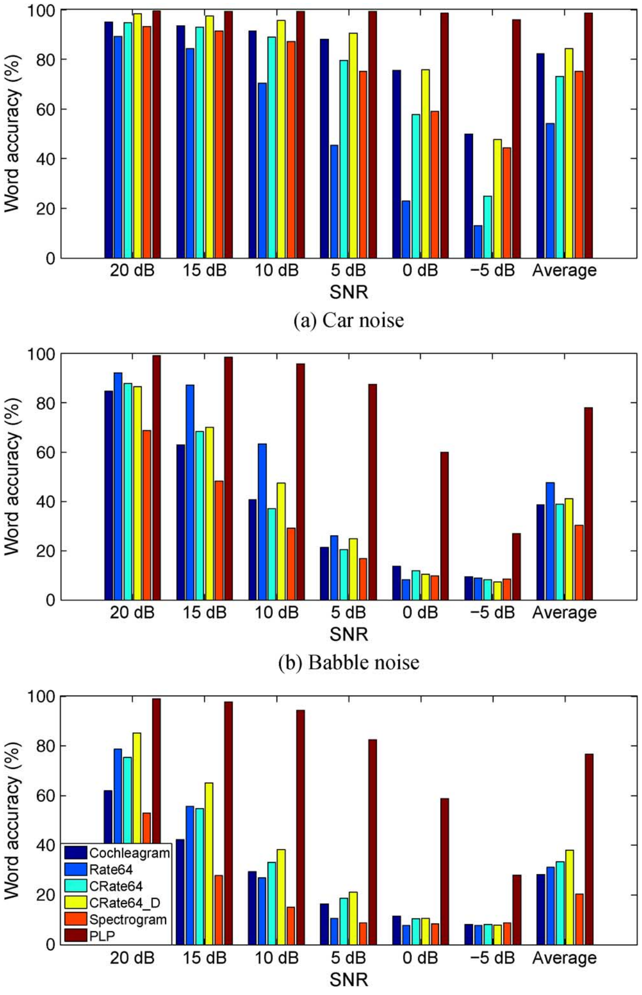
Feature	Feature Domain	Word Accuracy
Cochleagram	Spectral (non-linear frequency axis)	97.0
Rate64	Spectral (non-linear frequency axis)	93.2
CRate64	Spectral (non-linear frequency axis)	96.7
CRate64_D	Spectral (non-linear frequency axis)	98.7
Spectrogram	Spectral (linear frequency axis)	94.2
PLP	Cepstral	99.2

the silence model. The HMMs are trained in clean conditions using the HTK Toolkit [38]. Note that for marginalization-based ASR, HMMs are trained for each of the 5 features, whereas for the remaining two approaches they are trained using PLP cepstral features. The HTK decoder is adapted to perform bounded marginalization experiments. Additionally, word insertion penalties for each of the features and each of the methods are tuned separately using the development set of 240 test utterances.

1) Baseline Results: Before examining the performance of the various methods for incorporating the IBM in ASR, we first establish the baseline performance for each of the features previously described. Table I shows the word accuracy obtained in clean conditions. As expected, the best performance is obtained using PLP features as they reside in the cepstral domain as opposed to the other features that reside in the spectral domain. The next best performance is obtained using CRate64_D features. Rate64 performs the worst amongst the features that are considered.

In order to test robustness to additive noise, clean speech is mixed with three noise types from the NOISEX-92 corpus [39]—car noise, babble noise and factory noise, at 6 SNR conditions ranging from 20 dB to -5 dB, in decrements of 5 dB. Figs. 2(a)–2(c) show the word accuracy when trained HMMs are directly used to recognize noisy speech. Clearly, a marked deterioration in performance is observed for all features as the SNR decreases. Again, the best performance is obtained using PLP cepstral features. Notice that when additive noise is stationary (car noise), PLP features perform quite well, possibly because they are normalized and therefore, less affected by such noise types. Notice that for the other two noise types the decline in performance for the spectral features is quick and pronounced. In fact, the performance of the PLP features at 5 dB is comparable to or better than every other feature at greater SNRs.

2) IBM Results: Now that we have established the relevant baselines for our features, we examine the performance of the various methods for utilizing the IBM in ASR. Since marginalization results can be significantly affected by the feature used, we perform marginalization experiments using each of the previously described features to determine the best feature for comparison against other techniques. The marginalization results using the 5 spectral features are shown in Figs. 3(a)–3(c). Among the five features, CRate64_D performs the best in most conditions, likely due to the addition of the delta components. For babble noise at -5 dB, even though CRate64 performs slightly better than CRate64_D, the difference is not statistically significant at $p = 0.05$. At



(a) Car noise

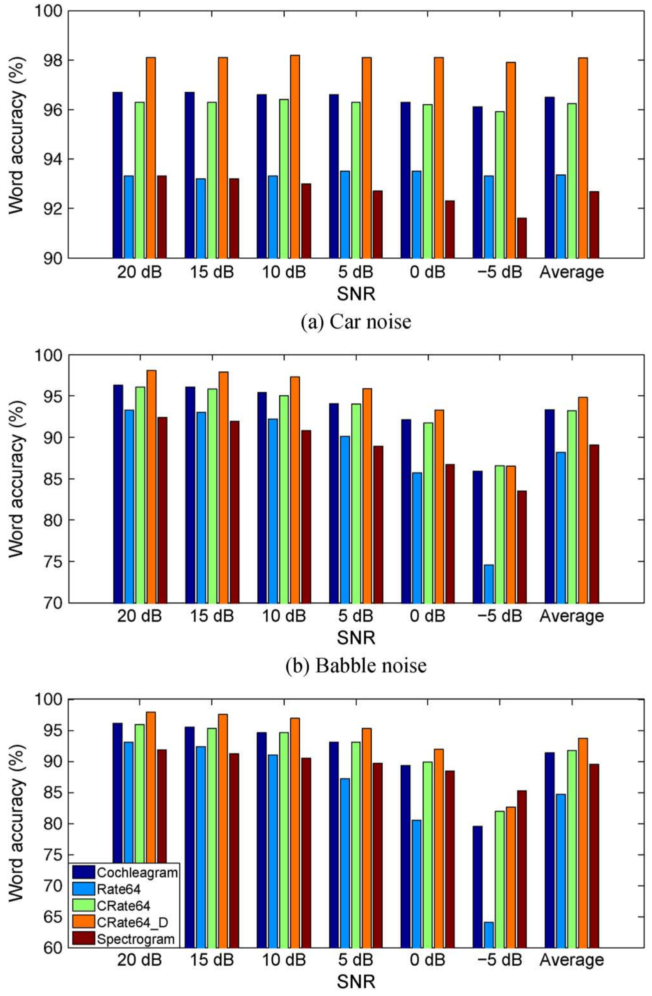
(b) Babble noise

(c) Factory noise

Fig. 2. Word accuracies in noisy conditions for 6 features and 6 SNR conditions from 20 dB to -5 dB, in decrements of 5 dB on TIDigits. Also shown is the average word accuracy for each feature, across all SNR conditions. (a) Car noise. (b) Babble noise. (c) Factory noise.

-5 dB the IBM is sparse and therefore, the delta mask is even sparser. Since delta components are fully marginalized during recognition, having a very sparse delta mask reduces the effect of adding delta components to the feature. Both cochleagram and CRate64 perform significantly better than Rate64 and obtain similar word accuracies in most conditions. Note that the rate of deterioration in performance with respect to SNR is lower for spectral features compared to the other features. But the peak performance of spectrogram (in clean conditions) is significantly lower than that of CRate64_D (see Table I). As a result, only for factory noise at-5 dB does it perform better than CRate64_D features. At high SNR conditions, it performs even worse than Rate64 features.

Next, we compare marginaliziation with the other two approaches—direct masking and reconstruction. The comparisons are presented in two parts, based on the domain in which marginalization and target speech segregation/reconstruction are performed. In the first part, they are performed using spectrogram features with a linear frequency axis. The results of this comparison are shown in Figs. 4(a)–4(c). In the second part, they are performed in the non-linear frequency domain. Since

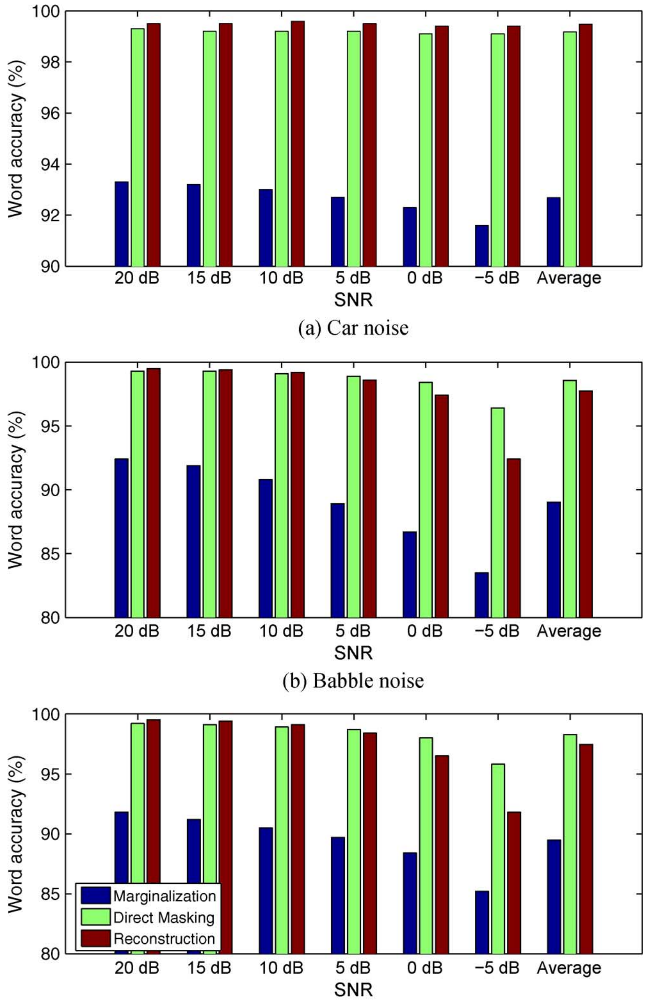


(a) Car noise

(b) Babble noise

(c) Factory noise

Fig. 3. Marginalization results for five spectral features in noisy conditions on TIDigits. Bounds are applied during marginalization in all the cases, except for the delta components of CRate64_D feature. Also shown is the average word accuracy across all SNR conditions. Note that the scale on the ordinate does not start at 0. (a) Car noise; (b) babble noise; (c) factory noise.



(a) Car noise

(b) Babble noise

(c) Factory noise

Fig. 4. Comparison of marginalization, direct masking, and reconstruction in the linear frequency domain on TIDigits. Marginalization uses spectral features. The other two approaches use PLP cepstral features. Also shown is the average word accuracy across all SNR conditions. (a) Car noise. (b) Babble noise. (c) Factory noise. Note the scale of the ordinate.

CRate64_D produced the best performance in this domain, it is chosen to represent marginalization. The corresponding results are shown in Figs. 5(a)–5(c).

As we can see from Fig. 4, when using the linear frequency domain, marginalization-based recognition performs significantly worse than both of the other approaches in all test cases. The performance gap between the direct masking approach and reconstruction is much closer. Only when the SNR drops below 5 dB on the two more difficult noise types, babble and factory, does the direct masking approach begin to outperform reconstruction. Since the IBM becomes very sparse in those cases, it is likely that the reconstruction suffers from the lack of reliable T-F units. It is unsurprising that performance at high SNRs and for car noise is comparable since baseline performance for PLP features in those cases was already strong.

The trends are somewhat different when using the non-linear frequency domain (see Fig. 5). The most obvious difference is the improvement seen for marginalization; in many cases its performance is now comparable to the other methods. Again, at low SNRs for the two more difficult noise types, the direct masking approach still performs better than reconstruction. However, in many cases the reconstruction does slightly outperform the direct masking approach. Our main point in comparing the two domains is to show that relative performance between direct masking and reconstruction is consistent. This allows us to connect these results to the results in the next section which focus on the linear frequency domain.

We would like to highlight a key point based on the results from this dataset. At no condition do the results justify a conclusion that the direct masking approach is not a viable method for incorporating the IBM in ASR. None of the results here provide strong evidence that the evaluated marginalization or reconstruction techniques are significantly better than the direct masking approach. In the next section we will examine whether similar conclusions hold on a larger dataset.

B. *AURORA4*

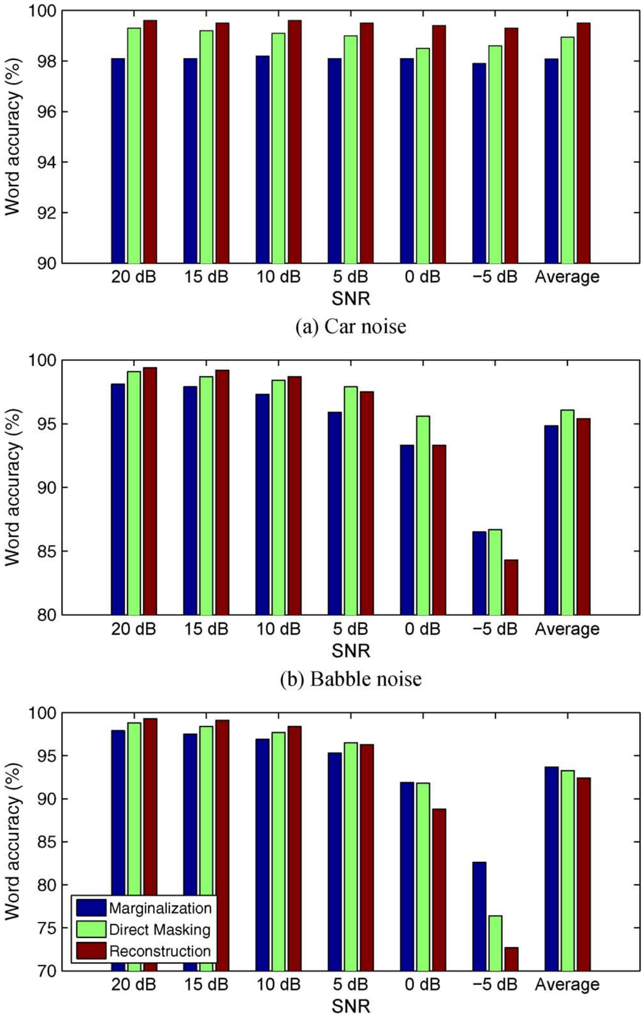
Our experimental setup here is very similar to the one used in the previous set of experiments. The Aurora4 [31] corpus is a 5000-word closed vocabulary task. It was generated by

TABLE II
WORD ACCURACY RESULTS USING THE IBM ON THE AURORA4 TEST SET. TIME DOMAIN REFERS TO FEATURES CALCULATED AFTER CONVERTING THE MASKED SIGNAL BACK TO THE TIME DOMAIN. SPECTRAL DOMAIN REFERS TO FEATURES CALCULATED DIRECTLY FROM THE MASKED SPECTRAL REPRESENTATION

System	Car	Babble	Restaurant	Street	Airport	Train	Average
Time Domain	86.3%	86.4%	86.2%	85.7%	87.4%	86.2%	86.4%
Spectral Domain	85.7%	87.1%	87.0%	85.6%	86.2%	84.9%	86.1%

TABLE III
WORD ACCURACY RESULTS USING THE IBM ON THE AURORA4 TEST SET. BASELINE IS THE UNSEGREGATED NOISY SPEECH

System	Car	Babble	Restaurant	Street	Airport	Train	Average
Baseline	72.7%	65.7%	63.3%	60.7%	65.0%	58.0%	64.2%
Reconstruction	84.3%	83.5%	84.1%	82.7%	84.5%	81.9%	83.5%
Direct Masking	86.3%	86.4%	86.2%	85.7%	87.4%	86.2%	86.4%
Perfect Reconstruction	90.2%	90.3%	90.2%	90.4%	90.7%	90.2%	90.3%



(a) Car noise

(b) Babble noise

(c) Factory noise

Fig. 5. Comparison of marginalization, direct masking, and reconstruction in the non-linear frequency domain on TIDigits. Also shown is the average word accuracy across all SNR conditions. Note that the scale on the ordinate does not start at 0. (a) Car noise; (b) babble noise; (c) factory noise.

adding noise to clean speech recordings in the Wall Street Journal (WSJ0) database [40]. Each utterance has been mixed with a noise source at a randomly chosen SNR between 5 and 15 dB. In total, six different noise types are used. Note that Aurora4 does not allow for a breakdown by SNR as each test set contains a mix of SNR conditions.

Using the HTK toolkit [38], we trained a baseline HMM recognizer on clean speech. Our models consisted of tied-state intra-word triphones with 16 Gaussians per state. The CMU dictionary was used for our pronunciations. Cepstral mean and variance normalized PLP features with delta and acceleration coefficients were used, giving a 39-dimensional feature vector. The reconstruction speech prior, consisting of a mixture of 1024 Gaussians, was also trained using the HTK. Again, the IBM was generated by comparing the premixed clean speech energy to the noise energy in the linear frequency domain using a local SNR threshold of 0 dB. After masking, the signal is converted back to the time domain before calculating features. As noted earlier, similar results are obtained by calculating features directly from the masked spectral representation as shown in Table II.

We performed recognition experiments to compare the use of masked and reconstructed speech.[3] Our results utilizing the IBM can be seen in Table III. Baseline refers to the recognition of unsegregated noisy speech. As expected, the addition of noise causes a significant drop in performance compared to word accuracy when recognizing clean speech, which is 91.7%. Reconstruction refers to speech where the masked regions have been estimated utilizing the technique described in Section III-B. When comparing these results to the baseline, we see a significant improvement. This is the type of comparison typically shown in the literature discussing spectral reconstruction [23], [27], [29]. With such improvements in accuracy over the baseline, it is easy to see how claims about the utility of reconstruction can be made.

However, these two results alone do not tell the whole story. Consider the direct masking results where no attempt to reconstruct masked units has been made; performance is better than reconstructed speech in every case. By attempting to reconstruct the missing spectral energy, performance was actually hindered. Combined with the results presented on TIDigits, this highlights a major issue within the missing-feature ASR literature. Without a comparison against the direct masking approach, it is unclear whether a particular reconstruction technique provides any benefit.

While our results show the direct masking approach significantly outperforms this particular reconstruction technique, we

[3] We did not perform marginalization-based experiments since the best performing spectral feature performed worse on clean speech than the baseline for any noise.

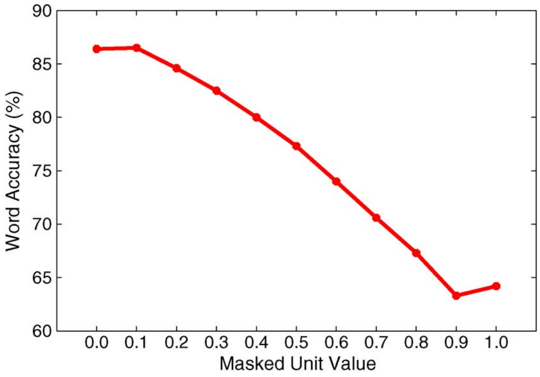


Fig. 6. Word accuracy results averaged over all noise conditions on the Aurora4 dataset. Results show the effect of manipulating the value for masked units with the ideal binary mask.

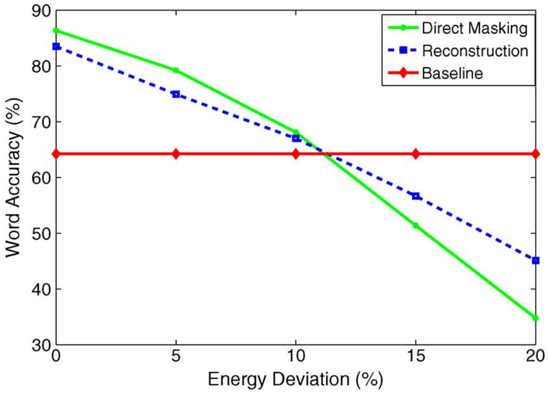


Fig. 7. Word accuracy results on the Aurora4 test set using randomly perturbed ideal binary masks. Average results over all 6 test conditions are displayed. Results on all test conditions follow the same general pattern. The results use an IBM where a given percentage of energy has been incorrectly classified as speech or noise dominant.

do not claim that the idea of reconstruction itself is ineffective. More sophisticated techniques can potentially surpass the simple direct masking approach. In Table III we also show results for perfect reconstruction, where every missing T-F unit has been replaced by the true energy of the clean speech. If the reconstruction worked perfectly, it would significantly outperform the direct masking approach.

As mentioned previously, the artificial zeros in the spectral value could potentially have unknown effects depending on the type of feature calculation software used. Instead of using a value of 0 for the masked units, any value between 0 and 1 can be substituted. Fig. 6 presents results for a range of values. Values near 0 produce nearly identical results, demonstrating that a small value near 0 can be safely substituted without decreasing performance. As the value used increases, performance does decrease, but this is to be expected; more noise energy is being included in the signal as the value increases.

We recognize that results using ideal masks may not tell the whole story since estimated masks do contain errors. We now examine the effects of mask errors on our results. We first examine the effects of randomly perturbing the ideal mask and then demonstrate results using estimated masks. To randomly perturb the masks a T-F unit is randomly and independently selected and its value is flipped; the process is repeated until a certain percentage of energy in the flipped T-F units has been reached. Obviously the errors introduced in this manner would differ from the errors seen in a mask estimation algorithm, but it does provide a general idea of the effect of mask errors and has been used in previous studies [17], [23].

Results on the Aurora4 test set are shown in Fig. 7. Average word accuracy across all 6 noise types for both the direct masking and reconstruction approaches versus energy deviation are shown. Energy deviation is the ratio of energy in the incorrectly labeled T-F units with respect to the total target energy. We use this metric as opposed to a simple count of unit labeling errors because we expect errors in high energy units to affect the final result more than those in low energy regions. Our results show reconstruction begins to outperform the direct masking approach at around 10% energy deviation on average. However, by the time reconstruction becomes the better performing metric, performance is similar to the baseline.

For our estimated mask experiments, we use a simple mask estimation technique. An estimate of noise and speech power is used to calculate the instantaneous SNR. Given the SNR estimate, a binary mask can be estimated. The noise estimate uses a power spectral density estimator recently proposed by Hendriks *et al.* [41] and the speech estimate comes from the work of Erkelens *et al.* [42]. In working with estimated masks, we learned that a small change to the direct masking approach can produce large performance improvements. Since the estimated mask contains errors, a less aggressive masking approach seems appropriate. Instead of using a binary mask with zeros for masked regions, any real value between 0 and 1 could be used. We found that a value of 0.5 performed best on a development set for this mask estimation algorithm. While this can be viewed as similar to soft-mask approaches, it still differs significantly; only a single uniform value is used for all masked units as opposed to estimating a specific value for each T-F unit. Results comparing the performance of the direct masking and reconstruction approach using estimated masks can be seen in Table IV. Both approaches provide a modest, but significant improvement over the baseline. As expected, based on our results using perturbed masks, there is no significant difference between the two approaches since the results are relatively close to the baseline. Fig. 8 shows the performance of direct masking using various values for the masked regions. Performance is robust to a range of values and values between 0.2 and 0.6 all perform well.

Results on ideal, estimated, and perturbed masks provide evidence that the simple direct masking approach does not perform significantly worse than reconstruction. In fact, just a small reduction in energy deviation would produce a larger increase in performance than perfect reconstruction would produce over the direct masking approach. Improvements to mask estimation may provide greater performance improvements compared to improvements to reconstruction methods.

The direct masking approach provides a stronger baseline compared to unsegregated noisy speech for missing feature ASR. We also believe this demonstrates that future work in mask estimation can evaluate performance in ASR without requiring more complicated reconstruction techniques. We

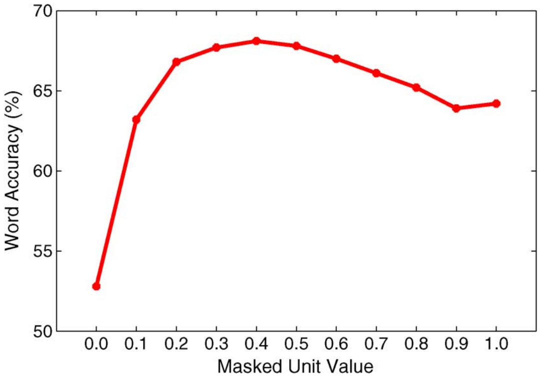


Fig. 8. Word accuracy results averaged over all noise conditions on the Aurora4 dataset. Results show the effect of manipulating the value for masked units in an estimated mask.

TABLE IV
Word Accuracy Results Using Estimated Binary Masks on the Aurora4 Test Set. Baseline is the Unsegregated Noisy Speech

System	Car	Babble	Restaurant	Street	Airport	Train	Average
Baseline	72.7%	65.7%	63.3%	60.7%	65.0%	58.0%	64.2%
Reconstruction	78.6%	67.0%	62.7%	65.4%	65.3%	66.1%	67.5%
Direct Masking	77.4%	68.3%	64.9%	65.8%	65.2%	65.1%	67.8%

acknowledge that the reconstruction approach evaluated in this study provides a relatively simple baseline. However, it still requires training general speech models and implementing the reconstruction technique. Direct masking operates without additional work. In the next section we will explain why our experiments showed, in contrast to previous results, that directly using binary-masked speech can work well in ASR.

V. Direct Masking and Variance Normalization

We have established that directly using the IBM can perform well, but why has this not been previously reported? The direct masking approach has been previously tested [17] and results were poor. Other studies likely ignored this approach based on these early results. If this is true, then what is different between our experimental setup and the likely setup of previous work? In our previous study [16], we found correlation between language model strength and recognition performance, suggesting that the Aurora results may have been due to the influence of the language model. However, the present study shows a similar effect for small vocabulary and large vocabulary tasks, indicating that the language model may not be a primary reason.

The remaining difference is the features used. Due to its popularity, previous work likely used MFCCs generated using the HTK. As already mentioned, our experiments used PLP features generated using the ICSI tool Feacalc [43]. In order to test our hypothesis that the feature type could drastically affect the results, we attempted to use the direct masking approach with MFCC features. Results on the TIDigits data mixed with factory noise are shown in Fig. 9. Performance for other noise types was similar. The direct masking approach using the IBM clearly does not work. In fact, it performs worse than no segregation at all. Obviously if previous researchers had seen a similar result, it would have served as a strong motivator to explore techniques for incorporating a binary mask in ASR.

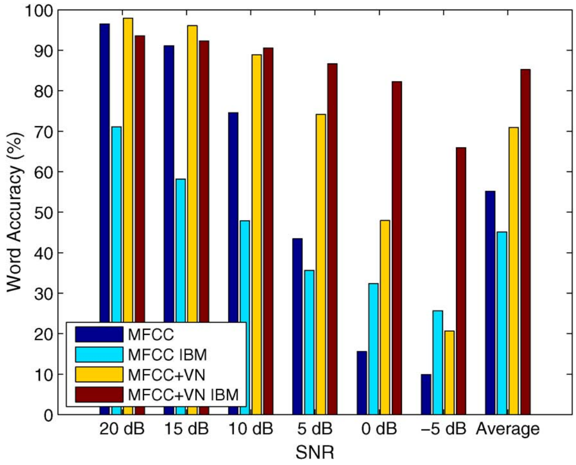


Fig. 9. Word accuracies for the factory noise condition on TiDigits for MFCCs with and without variance normalization and with and without the IBM. Results are shown for 6 SNR conditions from 20 dB to −5 dB, in decrements of 5 dB. Also shown is the average word accuracy across all SNR conditions.

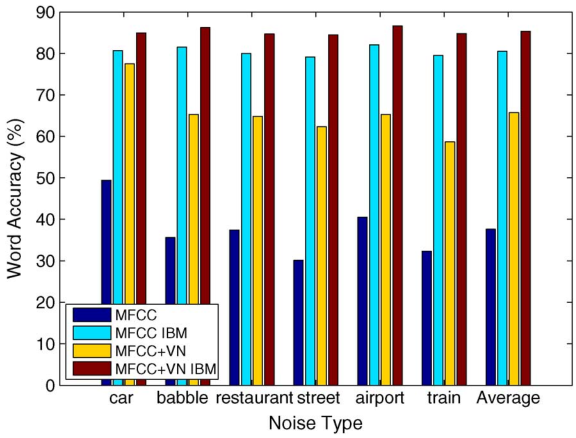


Fig. 10. Word accuracies for the Aurora4 dataset for MFCCs with and without variance normalization and with and without the IBM.

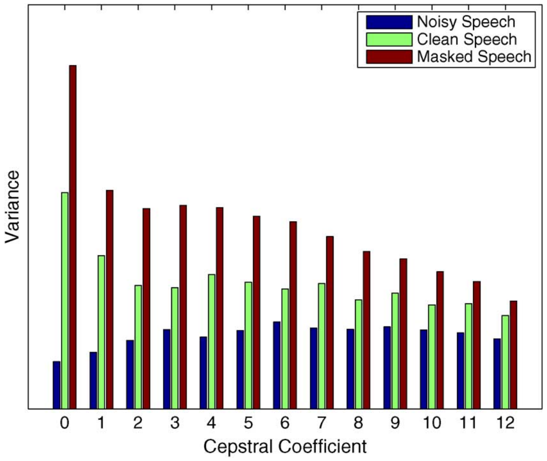


Fig. 11. Variances for the first 13 MFCCs computed from speech mixed with babble noise, clean speech, and IBM masked speech. The noise was mixed with speech from TIDigits at an SNR of 5 dB.

Many differences exist between the two feature types, but we found variance normalization was the only crucial difference. Although HTK-based features typically do not use variance normalization, HTK does provide this functionality. To show the effects of variance normalization, we perform it on the MFCC

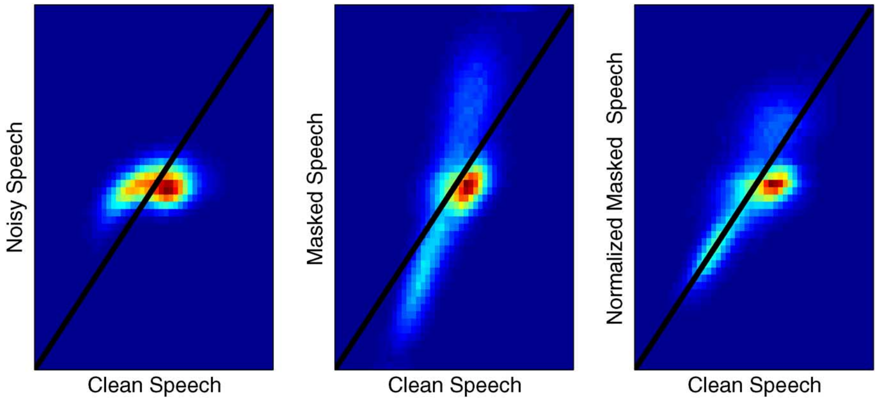


Fig. 12. Each plot shows the mean subtracted 3rd cepstral coefficient. The noisy speech has been mixed with factory noise at 5 dB SNR. The black line shows the plot of clean speech versus itself, illustrating the ideal relationship between enhanced speech and clean speech.

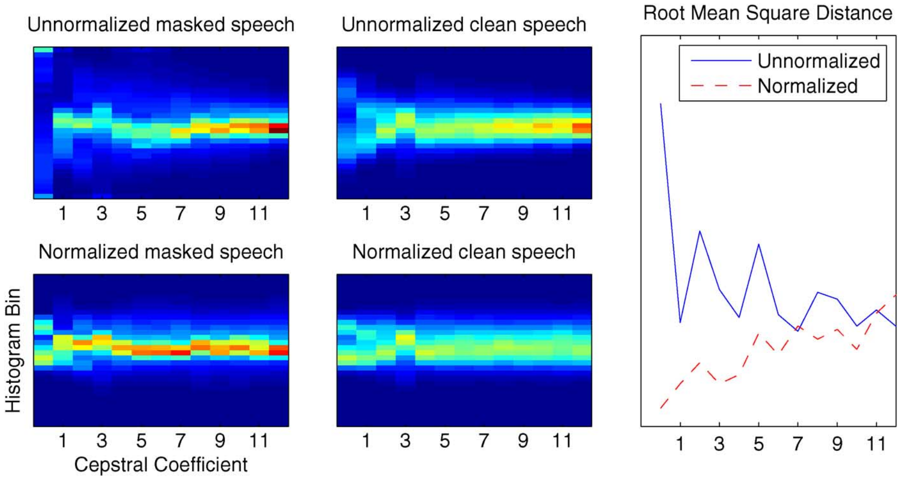


Fig. 13. Comparison of IBM masked MFCC features with clean speech. Noisy speech is created by mixing clean speech from TIDigits and factory noise at 5 dB SNR. Feature histograms in the top row are unnormalized while those in the bottom are normalized. The RMS distance on the right is computed over every point for each cepstral coefficient. The distance is between masked speech features and clean speech features.

features and show the results in Fig. 9. Each dimension was normalized to have a unit variance per utterance. Two things are immediately obvious when comparing the results in Fig. 9. First, variance normalization has improved every result. Even recognition on the noisy speech directly is significantly improved at lower SNRs. It appears the decreased variance in the features caused by the interference in the signal is a significant source of the performance degradation. Second, the direct masking approach now performs remarkably well. The benefits of using the IBM are only significant at lower SNRs. The variance normalized cepstral features themselves appear to be fairly robust in this small vocabulary domain. Regardless, simple variance normalization allows the direct use of the IBM to be a strong alternative to other techniques.

Fig. 10 presents the same experiments for Aurora4. On the larger vocabulary dataset, the binary masked MFCC features perform reasonably well even without variance normalization. As opposed to the TIDigits results, in all cases the binary masked features outperform the unenhanced features. The strength of the language model in this dataset may allow the recognizer to overcome the mismatch in the features. Even though the unnormalized features see an improvement on the Aurora4 dataset, variance normalization clearly provides a large boost in performance.

We can examine the average variance of the first 13 cepstral coefficients of the MFCC features used. Results are shown for babble noise at 5 dB SNR for the TIDigits data set in Fig. 11. The pattern is consistent across all noise types and SNRs. As expected, variances for features generated from noisy speech are less than features generated from clean speech. The noise effectively fills in T-F units with low energy and decreases the dynamic range of the clean speech. Variance normalization is commonly used in robust ASR; Chen and Bilmes provide a thorough analysis of the effects of normalization on noisy speech in [44].

Fig. 12 illustrates the effects of masking and variance normalization on the noisy speech for the third cepstral coefficient. In each plot, the noisy speech is shown versus clean speech. The data has been binned in order to show the density instead of the individual points. If the noisy speech matched the clean speech perfectly, only the bins along the diagonal marked by the black line would contain points. One can consider matching this ideal line as a goal for enhancement. The left subfigure shows that adding noise decreases the variance of the data. Noisy speech features also lack a linear relationship with clean speech. The middle subfigure shows that masking increases the variance and improves the linear relationship with clean speech, but the increased variance has caused the plot to deviate from the ideal line. Finally, the right subfigure demonstrates that variance normalization maintains the linear relationship and also corrects the deviation. Similar patterns are seen for other cepstral coefficients.

A more detailed look at the effects of variance normalization on masked speech over all cepstral coefficients can be seen in Fig. 13. The top row uses unnormalized features and the bottom row uses normalized features. The first column contains features calculated from IBM masked noisy speech and the second column contains features calculated from clean speech. In each plot, the data has been binned into histograms. When comparing the histograms of unnormalized clean speech and masked speech, the differences in the 0th feature are the most obvious. This is to be expected as both the addition of noise and the masking significantly change the amount of energy in any frame. Differences in other features are also more prominent than those seen in the normalized features. For the final column, the root mean square (RMS) distance of the corresponding points for each coefficient is shown. This plot uses the actual values for the features and is not a comparison of the histograms. The RMS plot clearly demonstrates that variance normalization significantly reduces the difference between masked and clean speech features, especially for lower order cepstral coefficients. We should note that, although Fig. 13 examines the 5 dB case, similar observations hold for other SNRs.

VI. Conclusion

We have shown the commonly held belief that a binary mask cannot be used directly in ASR is incorrect. In fact, directly using the IBM performs comparably to previously proposed missing data methods on a variety of datasets. Previous work likely missed this result due to the lack of variance normalization on acoustic features. By controlling the variance of the features, even results on the unsegregated noisy speech improved. Since the increase in variance appears to be a major issue, similar ASR systems should include variance normalization. We also demonstrated that nonzero mask values can significantly improve the performance of the direct masking approach when utilizing estimated masks.

The direct masking approach requires no additional overhead once a mask has been computed, and it is arguably the easiest approach to using a binary mask for ASR. While we certainly do not claim the direct masking approach should replace all missing data methods, we believe it presents a stronger baseline compared to unenhanced speech. Also, future work in speech enhancement may be able to evaluate their methods in terms of ASR performance without needing to implement more complicated missing feature methods.

References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, 1995.

[2] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[3] H. Hermansky, N. Morgan, and H.-G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," in *Proc. ICASSP*, 1993, vol. 10, pp. 509–512.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[5] A. S. Bregman*, Auditory Scene Analysis*. Cambridge, MA, USA: MIT Press, 1994.

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

[7] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 101–116, Sep. 2005.

[8] A. Narayanan and D. L. Wang, "The role of binary mask patterns in automatic speech recognition in background noise," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 8083–8093, 2013.

[9] M. Weintraub, "The GRASP sound separation system," in *Proc. IEEE ICASSP*, 1984, pp. 18A.6.1–18A.6.4.

[10] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[11] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[12] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary timefrequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.

[13] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[14] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, pp. 275–296, 2004.

[15] M. V. Segbroeck and H. V. Hamme, "Advances in missing feature techniques for robut large-vocabulary continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 1, pp. 123–137, Jan. 2011.

[16] W. Hartmann and E. Fosler-Lussier, "Investigations into the incorporation of the ideal binary mask in ASR," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4804–4807.

[17] M. Cooke, A. Morris, and P. Green, "Recognising occluded speech," in *Proc. ESCA Workshop Auditory Basis of Speech Percept.*, 1996, pp. 297–300.

[18] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proc. ICSLP*, 1994.

[19] S. Ahmad and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems 5 (NIPS'92)*, S. J. Hanson, J. D. Cowen, and C. L. Giles, Eds. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[20] R. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise," in *Proc. Eurospeech'97*, 1997, pp. 37–40.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[22] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, 2006.

[23] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2130–2140, Sep. 2007.

[24] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[25] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. Int. Conf. Spoken Lang.*, Beijing, China, 2000, pp. 373–376.

[26] J. V. Hout and A. Alwan, "A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4105–4108.

[27] W. Kim and J. H. L. Hansen, "Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in background noise conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2111–2120, Nov. 2010.

[28] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1984, pp. 111–114.

[29] J. Gemmeke and B. Cranen, "Noise reduction through compressed sensing," in *Proc. Interspeech*, 2008.

[30] Y. Wang and H. V. Hamme, "Multi-candidate missing data imputation for robust speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 17, 2012, doi:10.1186/1687-4722-2012-17.

[31] N. Parihar and J. Picone, "Analysis of the aurora large vocabulary extensions," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, vol. 4, pp. 337–340.

[32] J. Barker, M. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, pp. 5–25, 2005.

[33] S. Srinivasan and D. L. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Commun.*, vol. 52, pp. 72–81, 2010.

[34] , D. L. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York, NY, USA: Wiley-IEEE Press, 2006.

[35] J. Barker, M. Cooke, and D. P. Ellis, The RESPITE-CASA-Toolkit Project [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/J.Barker/ctk.html 2002

[36] S. Srinivasan, "Integrating computational auditory scene analysis and automatic speech recognition," Ph.D. dissertation, The Ohio State Univ., Columbus, OH, USA, 2006.

[37] M. Weintraub, "A theory and computational model of computational auditory scene analysis," Ph.D. dissertation, Stanford Univ., Stanford, NY, USA, 1985.

[38] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Cambridge Univ. Publishing Dept., 2002 [Online]. Available: http://htk.eng.cam.ac.uk

[39] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Speech Research Unit, Defense Research Agency, Malvern, UK, 1992, Tech. Rep..

[40] D. Paul and J. Baker, "The design of wall street journal-based CSR corpus," in *Proc. Int. Conf. Spoken Lang.*, Banff, AB, Canada, Oct. 1992, pp. 899–902.

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

[42] J. S. Erkelens, R. C. 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, no. 6, pp. 1741–1752, Aug. 2007.

[43] D. P. Ellis, J. A. Bilmes, E. Fosler-Lussier, H. Hermansky, D. Johnson, B. Kingsbury, and N. Morgan, "The SPRACHcore Software Package," [Online]. Available: http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html 2010

[44] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

William Hartmann (M'12) received the B.S. degree in computer science from Northern Kentucky University, in 2006, and the M.S. and Ph.D. degrees from the Ohio State University, in 2010 and 2012 respectively. His research interests include automatic speech recognition, natural language processing, and computational auditory scene analysis. Since 2012, he has been with LIMSI-CNRS, Paris.

Arun Narayanan (S'11) received the B.Tech. degree in computer science from the University of Kerala, Trivandrum, India, in 2005, and the M.S. degree in computer science from the Ohio State University, Columbus, USA, in 2012, where he is currently pursuing the Ph.D. degree. From November 2005 to June 2008, he was a System Engineer at IBM India.

His research interests include computational auditory scene analysis, robust automatic speech recognition, and machine learning.

Eric Fosler-Lussier (SM'05) received the B.A.S in Computer and Cognitive Studies and the B.A. in Linguistics from the University of Pennsylvania in 1993. He received the Ph.D. degree from the University of California, Berkeley in 1999; his Ph.D. research was conducted at the International Computer Science Institute. As an Associate Professor with the Department of Computer Science and Engineering (and Linguistics by courtesy) at The Ohio State University, he directs the Speech and Language Technologies (SLaTe) Laboratory. He currently serves on the IEEE Speech and Language Technical Committee, and received the 2010 Signal Processing Society Best Paper Award.

DeLiang Wang (F'04) photograph and biography not provided at the time of publication.