Nothing Doing: Re-evaluating Missing Feature ASR

William Hartmann, Arun Narayanan, Eric Fosler-Lussier, and DeLiang Wang

***Abstract*—Recently, much work has been devoted to the computation of binary masks for speech segregation. Common 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 is wrong; we demonstrate the effectiveness of directly using the masked data on both a small and large vocabulary dataset. In fact, this do-nothing approach outperforms two previously proposed 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 demonstrates that the field of missing feature ASR needs to be reconsidered.**

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 attempt to incorporate noise models into recognition [2]; these techniques typically require strong knowledge about the noise source and 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 making 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 refers to sound segregation based on the perceptual process of auditory scene analysis established 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

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

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. Figure 1 provides an example.

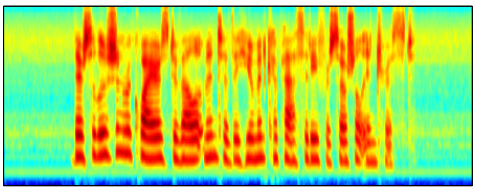
The IBM segregates the signal where a value of unity indicates that the corresponding T-F unit is grouped into the segregated target, and a value of zero indicates that the unit is considered part of the interference and hence removed [7]–[9]. We call T-F units with value 1 *unmasked*, and those with 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. [10]).

For ASR, common 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. Based on this common wisdom, many techniques have been proposed to incorporate the IBM or related binary masks in ASR. In this work, we will demonstrate that this common wisdom may, in fact, be a misconception. We build upon our previous results in [11], which showed the IBM could be used directly in ASR for large-vocabulary tasks, by examining results on a small-vocabulary dataset commonly used in early experiments. Our results show that directly using the IBM, which we term the *do-nothing* approach, not only works, but outperforms two main methods originally proposed to overcome the supposed inadequacies of the do-nothing approach. In addition, we explain the likely cause for the original misconception–the lack of variance normalization in the features used.

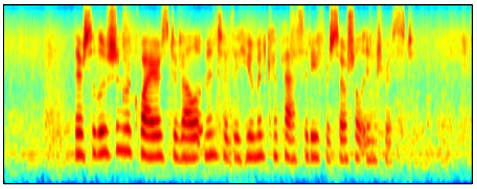
The rest of the paper is organized as follows. Section II presents background on the research incorporating the IBM in ASR. We describe in more detail two common approaches to

Technical Report OSU-CISRC-7/11-TR21

All four authors are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210; DLW is additionally with the Center for Cognitive Science. Contact email: hartmann.59@osu.edu



(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.

utilizing the IBM in ASR 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 explanation of 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, how to perform ASR on the segregated speech? Probably the first study to address this question is by Cooke et al. [12] 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. [13]) to perform HMM based ASR, where the key idea is to marginalize over missing or unreliable features in probability calculation (see also [14]). Early studies demonstrated the effectiveness of marginalizing missing data or features by removing feature components either in the spectral domain [12], [14] or in the cepstral domain [12], which is far superior to recognizing noisy speech without processing. Although not explicitly evaluated in those papers, direct recognition of present or reliable features without the marginalization step is expected to produce worse results. 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, later effort on missing feature recognition has focused on coupling with HMM recognizers using spectral features [15].

It is well known that, for ASR, spectral or T-F features are not nearly as effective as cepstral features [16]. 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 [17]. Treating marginalization as classifier compensation, Raj et al. [18] proposed 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. [19] 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 [17], [20].

Reconstruction of unreliable features from reliable ones is an inference that is inherently error-prone. It hence seems logical to ask the question: What if no reconstruction is attempted? To our surprise, *no single study* in the substantial literature on missing feature/data ASR has evaluated this simpler, do-nothing option, according to our systematic literature search. There are reasons to doubt the utility of the do-nothing approach. First, common wisdom would suggest that something must be done to holes (zeros) in a spectrogram or cochleagram created by binary masking [12]. This common 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 [18], [19], [21]. It is also possible that researchers obtained poor results with the do-nothing approach and declined to report the results. 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 fall under missing feature ASR. We separate the two techniques by how they handle missing features. For a more detailed review, see [22].

A. Marginalization-Based ASR

Originally proposed by Cooke et al., marginalization [15] was the first approach to address the issue of incorporating binary masks in ASR. While several variations were described

in [15], 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

$$p(X|Q_i) = \int p(X_r, X_u|Q_i)dX_u \tag{2}$$

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

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

where c is a particular Gaussian and M is the number of Gaussians in the GMM.

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 that X represents speech energy, then the true speech cannot have negative energy or more energy than in X_u. 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 [23], the marginalization approach performs remarkably well. However, performance on larger vocabulary systems degrades significantly [17], [18]. A likely cause is the use of spectral features instead of the cepstral features which are known to perform better in ASR [16]. 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. [18].

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 [18], [24], [25]. We will present a technique that has been previously shown to improve results over a baseline system [19]. A comparison between this method and directly using the IBM will allow us to determine if reconstruction is always the preferred approach.

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 $Y_r \cap Y_u = \emptyset$. 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 [18]. The speech prior, consisting of spectral features instead of the cepstral features eventually used for recognition, is modeled by a GMM. 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 [19]. 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. While this formulation can make use of a prior using full covariance matrices, our experiments, as in our previous work [11], use diagonal covariance matrices.

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. [21]), 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. We believe this assumption warrants examination and in the following section compare the performance of these techniques to using the mask directly.

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 directly applying the binary mask on both a small and large vocabulary dataset. The small vocabulary dataset, TIDigits [23], was used in Cooke et al. [15]. 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 [26].

A. TIDigits

The TIDigits corpus [23] consists of connected digit utterances. It has been widely used for speaker independent ASR studies [15], [27], [28]. 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. [15], we use the full test set to evaluate different ASR strategies. The do-nothing 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 [29]. To generate cochleagram based features, the signal is first passed through a 64 channel gammatone filterbank to perform T-F decomposition [29] (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 [15]. We use the CASA Toolkit [30] 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 [28]. Smoothed versions have also been used in other studies [27], [31].
- 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 [31], [32]. 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 [17]. 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 do-nothing 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 do-nothing 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 [29] (see Section 1.3.6), which is based on an approach introduced by Weintraub [33]. For the do-nothing 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 [29], [33].

In all three approaches, the IBM defined using a local SNR

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

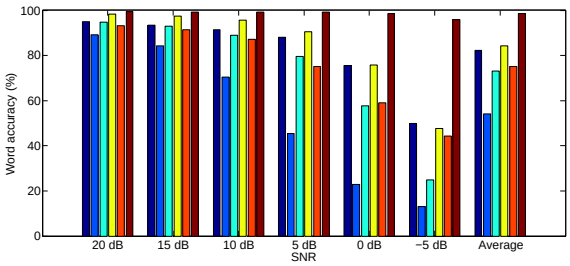
TABLE I
WORD ACCURACIES OBTAINED USING THE CLEAN TEST SET OF THE TIDIGITS CORPUS FOR VARIOUS FEATURES.

criterion of 0 dB is used to identify the masked and unmasked regions in the T-F representation of a noisy utterance [6].[1] 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 [31]. 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 do-nothing and reconstruction based strategies use the IBMs corresponding to the cochleagram and the spectrogram features.

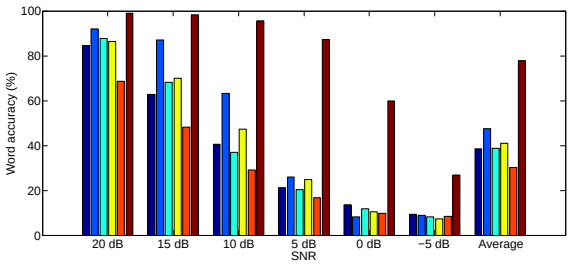
The ASR module consists of 11 word-level HMMs, one for each digit (1-9, 'oh' and 'zero') , a silence model, and a short-pause model. Each word-level HMM consists of 8 emitting states, with the observation probability modeled as a mixture of 10 diagonal Gaussian components [15], [19]. The short-pause model has only 1 state, tied to the middle state of the silence model. The HMMs are trained in clean conditions using the HTK Toolkit [34]. 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.

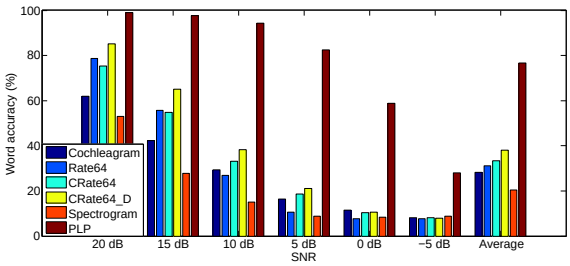
In order to test robustness to additive noise, clean speech is mixed with three noise types from the NOISEX-92 corpus [35] – car noise, babble noise and factory noise, at 6 SNR conditions ranging from 20 dB to -5 dB, in decrements of 5 dB. Figures 2(a) – 2(c) show the word accuracy when trained HMMs 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 5dB is comparable to or better than every other feature at greater SNRs.



(a) Car noise



(b) Babble noise

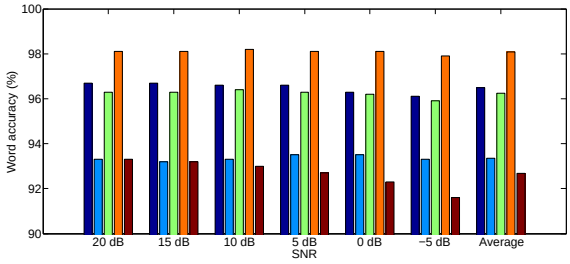


(c) Factory noise

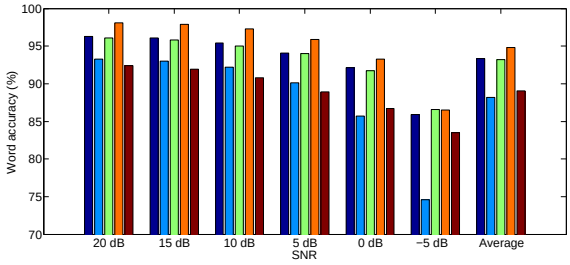
Fig. 2. Word accuracies in noisy conditions for 6 features and 6 SNR conditions from 20 dB to -5dB, in decrements of 5 dB. Also shown is the average word accuracy for each feature, across all SNR conditions. (a) Car noise. (b) Babble noise. (c) Factory noise.

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. The marginalization results using the 5 spectral features are shown in Figures 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

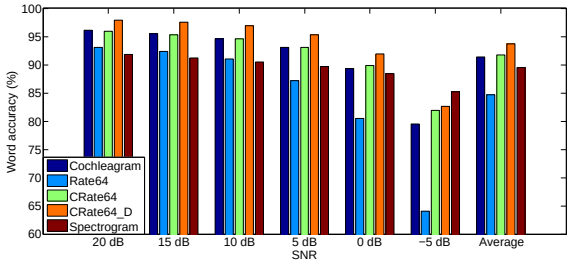
[1]Note that Cooke et al. [15] use a binary mask called the *a priori* mask based on whether the mixture energy is within 3 dB of the target energy, corresponding to a local criterion of 7.7 dB instead of 0 dB. We experimented with both the *a priori* mask and the IBM using a smaller development set of 240 utterances and found that the latter works better, and hence is used for marginalization-based ASR.



(a) Car noise

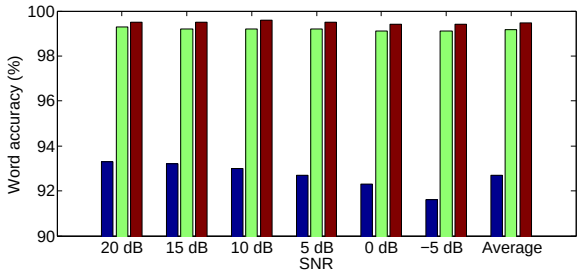


(b) Babble noise

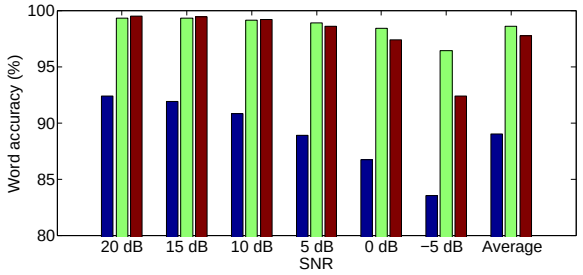


(c) Factory noise

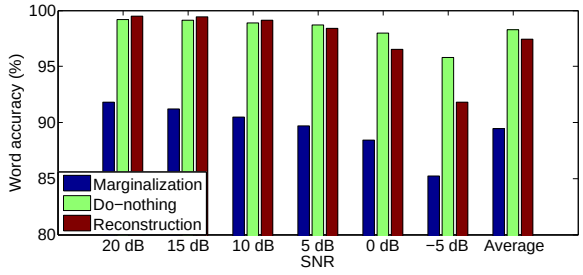
Fig. 3. Marginalization results for five spectral features in noisy conditions. Bounds are applied during marginalization in all the cases, except for the delta components of CRate64_D feature. Also shown is the average word accuracy across all SNR conditions. Note that the scale on the ordinate does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



(a) Car noise



(b) Babble noise



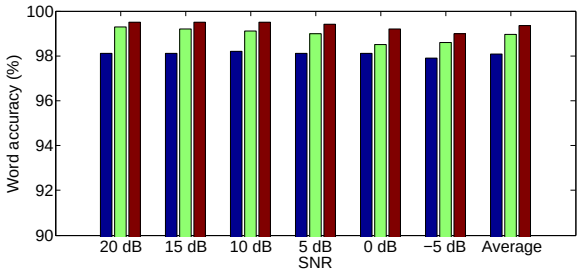
(c) Factory noise

Fig. 4. Comparison of marginalization, do-nothing, and reconstruction in the linear frequency domain. 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.

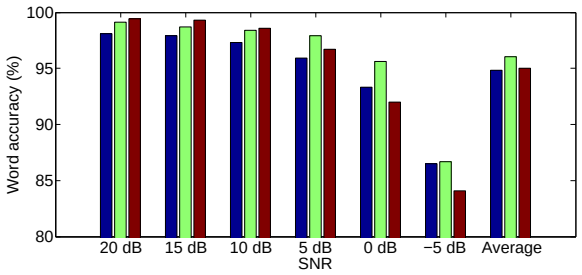
though CRate64 performs slightly better than CRate64_D, the difference is not statistically significant at $p = 0.05$. At -5 dB the IBM is sparse and therefore, the delta mask is even sparser. Since delta components are fully marginalized during recognition, having a very sparse delta mask reduces the 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 marginalziation with the other two approaches – do-nothing 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 Figures 4(a) – 4(c). In the second part, they are performed in the non-linear frequency domain. Since CRate64_D produced the best performance in this domain, it is chosen to represent marginalization. The corresponding results are shown in Figures 5(a) – 5(c).

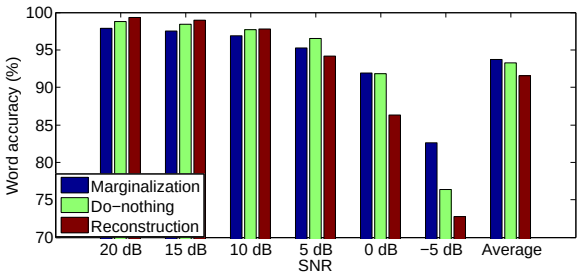
As we can see from Figure 4, when using the linear frequency domain, marginalization-based recognition performs significantly worse than both of the other approaches in all test cases. The performance gap between the do-nothing approach and reconstruction is much closer. Only when the SNR drops below 5dB on the two more difficult noise types, babble and factory, does the do-nothing 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



(a) Car noise



(b) Babble noise



(c) Factory noise

Fig. 5. Comparison of marginalization, do-nothing, and reconstruction in the non-linear frequency domain. Also shown is the average word accuracy across all SNR conditions. Note that the scale on the ordinate does not start at 0.

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 Figure 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 do-nothing approach still performs better than reconstruction. However, in many cases the reconstruction does slightly outperform the do-nothing approach.

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 do-nothing approach is not a viable method for incorporating the IBM in ASR. None of the results here provide strong evidence that marginalization or reconstruction are significantly better than the do-nothing 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 [26] corpus is a 5000-word closed vocabulary task. It was generated by adding noise to clean speech recordings in the Wall Street Journal (WSJ0) database [36]. Each utterance has been mixed with a noise source at a randomly chosen SNR between 5 and 15dB. In total, six different noise types are used.

Using the HTK toolkit [34], we trained a baseline HMM recognizer on clean speech. Our models consisted of tied-state inter-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 0dB.

We performed recognition experiments to compare the use of masked and reconstructed speech.[2] Our results utilizing the IBM can be seen in Table II. 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 [19], [21], [24]. 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 do-nothing results where no attempt to reconstruct masked units has been made. Its 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 with in the missing-feature ASR literature. Without a comparison against the do-nothing approach, it is unclear whether a particular reconstruction technique provides any benefit.

While our results show the do-nothing approach significantly outperforms this particular reconstruction technique, we do not claim that the idea of reconstruction itself is ineffective. More sophisticated techniques could potentially surpass the simple do-nothing approach. In Table II 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

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

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%
Do-nothing	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%

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

reconstruction worked perfectly, it would significantly outperform the do-nothing approach. Nonetheless, future studies should utilize the do-nothing approach as a baseline rather than unsegregated noisy speech. In the next section we will explain why our experiments showed, in contrast to previously held beliefs, that directly using binary-masked speech can work well in ASR.

V. WHY IS DO-NOTHING IGNORED?

We have established that directly using the IBM can perform well, but why has this been missed by previous researchers? We believe that this do-nothing approach was most likely tested previously, but the results were poor and went unreported. If this is true, then what is different between our experimental setup and the likely setup of previous work? In our previous study [11], 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 [37]. In order to test our hypothesis that the feature type could drastically affect the results, we attempted to use the do-nothing approach with MFCC features. Results on a 240 utterance subset of TIDigits are shown in Figure 6. The do-nothing 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.

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, it is a commonly used technique in the field. To show the effects of variance normalization, we perform it on the MFCC features and show the results in Figure 6. Each dimension was normalized to have a unit variance per utterance. Two things are immediately obvious when comparing the results in Figure 6. First, variance normalization has improved every result. Even recognition on the noisy speech directly is significantly improved at lower SNRs. It appears the increased variance in the features caused by the interference in the signal is a significant source of the performance degradation. Second, the do-nothing approach

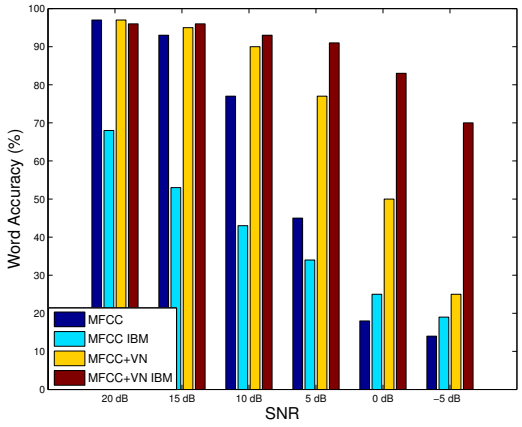


Fig. 6. Comparison of features with and without variance normalization (VN). Results are on a 240 utterance subset of TIDigits mixed with factory noise. MFCC refers to features generated from the unenhanced noisy speech while MFCC IBM refers to features generated from the IBM masked speech.

now performs remarkably well. The simple use of variance normalization allows the direct use of the IBM to be a strong alternative to other techniques.

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 outperforms more complicated techniques 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. It is possible that speech enhancement methods simply mitigate this issue.

While much research has been done in missing feature ASR, it may be built on a faulty foundation. We believe the initial work in marginalization and reconstruction strongly influenced the focus of subsequent work. With the success of missing feature ASR compared to direct recognition of noisy speech and the drive to improve existing techniques, it may be difficult to see a re-evaluation like the one presented here. We also hope this work serves as a lesson: with so many pieces interacting with one another in a typical ASR system, a small change in one part can make significant differences in the end result. Here, a simple variance normalization to the features used produced drastic changes in the performance of the do-nothing

approach. Hence healthy skepticism on common practice is called for.

ACKNOWLEDGMENT

The work of WH and EFL was supported in part by an NSF CAREER grant (ISI-0643901). The work of AN and DLW was supported in part by an AFOSR grant (FA9550-08-1-0155).

REFERENCES

[1] G. Yifan, "Speech Recognition in Noisy Environments: A Survey," *Speech Communications*, vol. 16, pp. 261–291, 1995.

[2] M. J. F. Gales and S. J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352–359, 1996.

[3] H. Hermansky, N. Morgan, and H.-G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing," in *Proceedings of ICASSP*, vol. 10, 1993, pp. 509–512.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[5] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, Mass: 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: Kluwer Academic, 2005, pp. 181–197.

[7] M. Weintraub, "The GRASP sound separation system," in *Proceedings of IEEE ICASSP*, 1984, pp. 18A.6.1–4.

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

[9] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, pp. 684–697, 1999.

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

[11] W. Hartmann and E. Fosler-Lussier, "Investigations into the incorporation of the ideal binary mask in asr," in *Proceedings of IEEE ICASSP*, Prague, Czech Republic, May 2011.

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

[13] 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: Morgan Kaufmann, 1993.

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

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

[16] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, 1980.

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

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

[19] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2130–2140, 2007.

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

[21] W. Kim and J. H. L. Hansen, "Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in basckground noise conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2111–2120, November 2010.

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

[23] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proceedings IEEE International Conference on Speech Acoustics and Signal Processing*, 1984, pp. 111–114.

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

[25] M. V. Segbroeck and H. V. Hamme, "Advances in missing feature techniques for robut large-vocabulary continuous speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 19, no. 1, pp. 123–137, January 2011.

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

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

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

[29] D. L. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

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

[31] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proceedings of International Conference on Spoken Language*, Beijing, China, 2000, pp. 373–376.

[32] S. Srinivasan, "Integrating computational auditory scene analysis and automatic speech recognition," Ph.D. dissertation, The Ohio State University, 2006.

[33] M. Weintraub, "A theory and computational model of computational auditory scene analysis," Ph.D. dissertation, Stanford University, 1985.

[34] 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 University Publishing Department, 2002. [Online]. Available: http://htk.eng.cam.ac.uk

[35] 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, Tech. Rep., 1992.

[36] D. Paul and J. Baker, "The design of wall street journal-based CSR corpus," in *Proceedings of International Conference on Spoken Language*, Banff, Alberta, Canada, October 1992, pp. 899–902.

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