INVESTIGATIONS INTO THE INCORPORATION OF THE IDEAL BINARY MASK IN ASR

William Hartmann and Eric Fosler-Lussier

The Ohio State University
Department of Computer Science and Engineering
{hartmanw, fosler}@cse.ohio-state.edu

ABSTRACT

While much work has been dedicated to exploring how best to incorporate the Ideal Binary Mask (IBM) in automatic speech recognition (ASR) for noisy signals, we demonstrate that the simple use of masked speech can outperform standard spectral reconstruction methods. We explore the effects of both the accuracy of the mask estimation and the strength of the language model on our results. The relative performance of these techniques is directly tied to the accuracy of the estimated mask. Although the use of masked speech fails when significant numbers of errors are present, the maximum performance for spectral reconstruction techniques also drops significantly. This implies improvements in mask estimation can provide greater gains in ASR performance than improvements in the incorporation of the IBM in ASR. Previous work may have ignored the direct use of masked speech due to its poor performance on tasks without a strong language model.

Index Terms— robust automatic speech recognition, spectral reconstruction, ideal binary mask

1. INTRODUCTION

The IBM has been proposed as a computational goal for speech enhancement [1]. In the time-frequency domain, each time-frequency (T-F) unit is classified as either speech or noise dominant. The IBM masks all noise-dominant regions and only leaves energy in the T-F units which are speech-dominant; speech and noise dominance is determined by the local SNR at each T-F unit. While this knowledge is assumed to be known a priori for the IBM, many techniques exist to estimate the IBM. Perceptual studies have demonstrated that the IBM can improve speech intelligibility [2].

Unlike the use of the IBM for speech intelligibility, incorporation of the IBM in ASR has been assumed to be a difficult problem. This assumption stems from how the cepstral features used in ASR are calculated; essentially, the cepstral features are weighted sums of the T-F units. Since the IBM sets any noise-dominant T-F unit to zero, its true energy is not included in the summation. Common wisdom in the field holds that leaving these missing values as zero will negatively impact feature calculation and subsequent ASR performance.

One of the first techniques developed to incorporate the IBM in ASR is the missing data recognizer [3], a modified HMM-based recognizer. By considering the speech dominant units reliable and the noise dominant units unreliable, it decodes an utterance by evaluating the reliable units and integrating over the unreliable units. This allows the unreliable units to be treated as what they truly are, missing. While this approach performs quite well on small-vocabulary tasks, performance degrades with larger vocabularies [4]. One of the main issues is that the system forces the use of spectral features which do not perform as well as cepstral features.

An alternative technique, spectral reconstruction, builds on ideas from missing data recognition. Instead of treating the unreliable units as missing, they are estimated based on the information obtained from the reliable units [4]. Several techniques have been proposed which have produced promising results, such as the prior-based method of Raj et al. [4] and the exemplar-based method of Gemmeke and Cranen [5]. Since this method reconstructs the missing spectral information, cepstral features can now be computed. However, these these aforementioned techniques ignore one simple alternative. Instead of reconstructing the speech, we can use the masked speech, where any T-F unit masked by the IBM is treated as zero, directly in our feature calculation. Surprisingly, we find that this significantly outperforms the use of reconstructed speech. It demonstrates that spectral reconstruction is a difficult task and comparisons against unenhanced baseline systems are not adequate since a poor reconstruction can actually decrease ASR accuracy compared to masked speech. We discuss spectral reconstruction in more detail, including the specific technique we use, in Section 2. Section 3 presents our experimental setup. In Sections 4 and 5 we present our results and conclusions respectively.

2. SPECTRAL RECONSTRUCTION

Spectral reconstruction refers to the general technique of estimating the unreliable or missing T-F units. Several techniques have been proposed, but we will focus on the prior-based method as introduced in [4]. Using a binary mask, the noisy speech vector Y is partitioned 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 this noisy spectral vector Y, we want to estimate the true spectral vector $\hat{X}$ for the clean speech.

In order to estimate $\hat{X}_u$, a speech prior is used. The speech prior, consisting of spectral features instead of the cepstral features eventually used for recognition, is modelled 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.

$$\mu_c = \begin{bmatrix} \mu_{r,c} \\ \mu_{u,c} \end{bmatrix} \quad \Sigma_k = \begin{bmatrix} \Sigma_{rr,c} & \Sigma_{ru,c} \\ \Sigma_{ur,c} & \Sigma_{uu,c} \end{bmatrix} \tag{1}$$

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

$$\hat{X}_u = \sum_{c=1}^{M} p(c|X_r)\hat{X}_{u,c} \tag{2}$$

where M is the number of mixtures and $\hat{X}_{u,c}$ is the expected value of X given the cth mixture. To estimate $p(c|X_r)$, the marginal distribution $p(X_r|c) = N(X_r; \mu_{r,k}, \Sigma_{rr,k})$ is used [6]. Finally, we compute the expected value of X_u given the cth mixture by

$$\hat{X}_{u,c} = \mu_{u,c} + \Sigma_{ur,c}\Sigma_{rr,c}^{-1}(X_r - \mu_{r,c}). \tag{3}$$

The unreliable portion of the spectral vector is then replaced by the estimate $\hat{X}_u$ and cepstral features can be computed from the reconstructed spectrogram. While this formulation can make use of a prior using full covariance matrices, our experiments, as in [6], use diagonal covariance matrices.

3. EXPERIMENTAL SETUP

Using the HTK [7] system, we trained a baseline recognizer on clean speech from the Aurora4, 5000-word closed vocabulary task. This task is a modification of the Wall Street Journal (WSJ0) database where noise has been added to the clean speech recordings. Each utterance has been mixed at a randomly chosen SNR between 5 and 15dB. The CMU dictionary was used for our baseline pronunciations. Tied-state inter-word triphones with 16 Gaussians per state comprised the acoustic model using 39 phones. Mean-subtracted PLP features with delta and double-delta coefficients were used, giving a 39-dimensional feature vector. The reconstruction speech prior, consisting of a mixture of 1024 Gaussians, was also trained using HTK.

The IBM was generated by comparing the speech and noise components of the noisy speech to determine the true instantaneous SNR at each T-F unit. Any T-F unit with an SNR greater than 0dB was considered speech-dominant and remaining units were masked. The estimated binary mask was generated similarly except the instantaneous SNR was estimated via the Log-MMSE algorithm as implemented in [8].

4. RESULTS

We performed several recognition experiments to compare the use of masked and reconstructed speech. Our initial results utilizing the IBM can be seen in Table 1. Baseline refers to the recognition of unenhanced noisy speech. The addition of noise causes a significant drop in performance compared to word accuracy when recognizing clean speech. Reconstructed speech refers to speech where the masked regions have been estimated utilizing the technique described in Section 2. When comparing these results to the baseline, we see a significant improvement. This is the type of comparison typically shown in papers discussing spectral reconstruction [5, 6]. With such great improvements in accuracy over the baseline, it is easy to see how claims about the quality of reconstruction can be made. However, when considering the masked speech results, it becomes evident that reconstruction has actually hindered performance compared to simply treating masked regions as having zero energy. This surprising result demonstrates that improvement over an unenhanced baseline may not demonstrate the efficacy of a reconstruction technique. However, we do not claim that reconstruction cannot provide improvements over simply using masked speech. The perfect reconstruction result, which shows performance when the masked T-F units have been replaced by clean speech, demonstrates a ceiling for performance using reconstruction that is significantly better than masked speech.

Early systems using both missing data recognition and spectral reconstruction generally reported results on small vocabulary tasks where the language model would have provided little to no help. Perhaps the use of masked speech was originally discarded due to its poor performance on small vocabulary tasks. In order to examine this phenomenon, we performed experiments using reduced language models. Table 2 shows results for a unigram language model and Table 3 shows results for a uniform language model. The unigram results still show a significant improvement for both masked and reconstructed speech over baseline results. In contrast, this improvement disappears when using a uniform language model. While masked speech still outperforms reconstructed speech, neither produces great improvements over baseline results. Both methods were unable to cope without the benefits that a language model provides.

While the IBM is a goal, it has to be estimated in practice. An estimated mask will contain errors and may affect the effectiveness of masked speech versus reconstructed speech. We examine the effects of an imperfect mask by using the Log-MMSE algorithm to estimate the binary mask. Recognition results for the estimated mask are shown in Table 4. It should be noted that overall performance might be worse than expected because we used the algorithm only for mask estimation; we did not use the algorithm to enhance the speech. Our results show that in the case of an estimated binary mask, reconstructing the speech performs much better than simply

System	car	babble	restaurant	street	airport	train
Baseline	72.7%	65.7%	63.3%	60.7%	65.0%	58.0%
Reconstructed Speech	84.3%	83.5%	84.1%	82.7%	84.5%	81.9%
Masked Speech	86.3%	86.4%	86.2%	85.7%	87.4%	86.2%
Perfect Reconstruction	90.2%	90.3%	90.2%	90.4%	90.7%	90.2%

Table 1. Word accuracy results using IBM on the Aurora4 test set. Baseline is the unenhanced noisy speech. Reconstructed speech is masked speech where the masked T-F units have been estimated using spectral reconstruction. Masked speech has zero energy in the masked T-F units. Perfect reconstruction is the accuracy of a system where all masked T-F units have been replaced with clean speech. Accuracy on clean speech is 91.7%. Masked speech significantly outperforms reconstructed speech in all test cases.

System	car	babble	restaurant	street	airport	train
Baseline	46.1%	41.1%	40.6%	35.7%	39.5%	35.1%
Reconstructed Speech	57.4%	55.5%	56.6%	54.9%	56.9%	53.8%
Masked Speech	57.6%	58.9%	58.3%	56.7%	58.8%	56.9%
Perfect Reconstruction	64.2%	64.1%	64.2%	64.0%	64.3%	63.7%

Table 2. Word accuracy results using IBM using a unigram language model. While all results have dropped, both techniques still outperform baseline results.

System	car	babble	restaurant	street	airport	train
Baseline	19.4%	23.4%	30.7%	28.7%	25.1%	30.6%
Reconstructed Speech	17.0%	15.4%	17.0%	10.1%	18.2%	10.3%
Masked Speech	22.4%	25.3%	25.1%	21.1%	26.1%	22.6%
Perfect Reconstruction	38.8%	39.4%	40.0%	38.6%	40.7%	39.9%

Table 3. Word accuracy results using IBM with a uniform language model. All results have dropped and neither technique outperforms the baseline system.

System	car	babble	restaurant	street	airport	train
Reconstructed Speech	74.9%	65.4%	57.8%	58.2%	63.1%	61.4%
Masked Speech	60.5%	44.7%	39.4%	42.0%	44.3%	47.9%
Perfect Reconstruction	86.7%	76.1%	69.0%	76.3%	70.9%	81.5%

Table 4. Word accuracy results using estimated binary mask. The estimated mask contains a significant amount of errors and the masked speech no longer outperforms reconstructed speech. Perfect reconstruction has also greatly suffers from mask estimation errors.

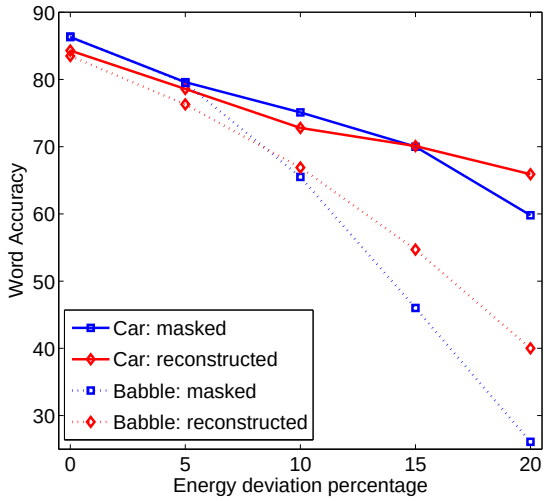

Fig. 1. Effects of mask errors randomly added to the IBM on word accuracy for two representative noise cases.

masking the speech. However, the perfect reconstruction results also suffer significantly. Even the optimal reconstruction cannot fully recover from a poorly estimated mask.

Given this result, the obvious question is how much must the mask deviate from the IBM before the significant improvements gained from using masked speech disappears. In Figure 1, we plot how the recognition accuracy changes due to errors in the estimated mask. We simulate estimation error by randomly perturbing the IBM. We measure the deviation from the IBM as the percentage of energy changed [6]. Total energy is the overall energy present in the speech masked by the IBM and deviated energy is the amount of energy in the areas where the mask has been changed. The deviation percentage, the x-axis, is the ratio of deviated energy to total energy. While the errors in these artificially created masks do not resemble the errors made by a real estimation algorithm, it still allows the robustness of both masked speech and reconstructed speech to mask estimation errors to be examined. Two representative noise types, car and babble, were examined. As the percentage of deviation increased, recognition accuracy drops for all cases. However, the performance of masked speech drops quicker and reconstructed speech begins to outperform masked speech.

5. CONCLUSION

The simple use of masked speech can outperform current spectral reconstruction algorithms when using an IBM. The performance of both masked and reconstructed speech is tied to the quality of the estimated mask. As the quality of the mask degrades, reconstructed speech begins to outperform masked speech. However, the ceiling for reconstructed speech also greatly degrades. Improvements in mask estimation could provide much greater recognition gains than improvements in reconstruction techniques. The use of masked speech may have previously been ignored due to its poor performance without the aid of a strong language model. While these results do not eliminate the usefullness of reconstruction techniques, it does demonstrate that a technique must present stronger evidence for its efficacy than simply comparing performance to an unenhanced baseline. In the future we would like to explore these results using more state of the art mask estimation algorithms. We would also like to examine if uncertainty decoding [9] can produce similar improvements for masked speech as it does for reconstructed speech [6].

Acknowledgements. The authors would like to thank DeLiang Wang, Arun Narayanan, and Xiaojia Zhao for their helpful comments. This work was supported in part by NSF grant IIS-0643901.

6. REFERENCES

[1] 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., pp. 181–197. Kluwer Academic, Norwell MA, 2005.

[2] N. Li and P. C. Loizou, "Factors influencing intelligibility of binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, pp. 1673–1682, 2008.

[3] 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.

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

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

[6] 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.

[7] 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.

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

[9] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 412–421, 2005.