ASR-DRIVEN TOP-DOWN BINARY MASK ESTIMATION USING SPECTRAL PRIORS

William Hartmann and Eric Fosler-Lussier

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

ABSTRACT

Typical mask estimation algorithms use low-level features to estimate the interfering noise or instantaneous SNR. We propose a simple top-down approach to mask estimation. The estimated mask is based on a specific hypothesis of the underlying speech without using information about the interference or the instantaneous SNR. In this pilot study, we observe a 9% reduction in word error over a baseline recognition system on the Aurora4 corpus, though much greater gains could theoretically be achieved through improvements to the model selection process. We also present SNR improvement results showing our method performs as well as a standard MMSE-based method, demonstrating that speech recognition can aid speech enhancement. Thus, the relationship between recognition and enhancement need not be one way: linguistic information can play a significant role in speech enhancement.

Index Terms— robust automatic speech recognition, ideal binary mask, mask estimation

1. INTRODUCTION

One of the longstanding issues in automatic speech recognition (ASR) is the acoustic mismatch between training and testing data caused by the presence of noise. Many methods have been proposed to increase the robustness of ASR systems to this problem. One approach uses the ideal binary mask (IBM) to improve signal quality [1]. Given a spectral representation of the input signal, the local SNR for each time-frequency (T-F) unit can be calculated. The mask has a value of unity where the local SNR exceeds some threshold and zero elsewhere. More formally, the IBM is defined as

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

where f is a frequency band and t represents a particular time frame. $S(f,t)$ and $N(f,t)$ represent the amount of energy at a T-F unit for the clean speech and interfering noise respectively. The threshold θ is typically set to 0.

Once the mask has been calculated, it is multiplied by the original signal to suppress noise-dominant T-F regions. The IBM relies on *a priori* knowledge of the separate speech and interference signals. In practice, only the mixed signal is observed, so the mask must be estimated. Typical estimation methods use low level signal information in attempting to estimate the interfering noise. In this pilot study, we propose an alternative approach that uses top-down linguistic information to estimate the characteristics of the underlying speech signal and is largely agnostic to the interference present in the signal. We show that recognition results can aid mask estimation, and ultimately, speech enhancement.

In Section 2, we outline previous approaches to mask estimation, including other top-down approaches. Our mask estimation approach is presented in Section 3. Specific details of our experimental setup are described in Section 4 and in Section 5 results for both ASR and SNR improvements are discussed. Section 6 presents conclusions and directions for future work.

2. PREVIOUS WORK IN MASK ESTIMATION

Many mask estimation algorithms aim to estimate the noise in Equation 1. One standard method of estimating the noise is spectral subtraction, as used in [2]. Here, the noise is modeled by averaging the non-speech frames in the signal. The instantaneous SNR is then calculated and used to determine the binary mask classification. Improved methods have also been developed that increase the accuracy of the noise estimation by tracking changes in non-stationary noise sources [3]. The focus of these techniques is modeling the noise source.

In the spectral subtraction-based techniques, the estimated clean speech is obtained by subtracting the estimated noise from the noisy speech signal. More recent techniques have looked at estimating both the noise and the speech. In [4], instead of comparing the estimated clean speech and noise signals, models of the clean speech and noise-corrupted speech are compared. Here, the focus is on the speech and how the noise impacts the clean speech characteristics. While this system produces an improvement over standard spectral subtraction-based methods by using speech models, it does not use higher level linguistic information.

One of the first systems to couple the recognition and mask estimation process was that of Barker et al. [5], where

the mask and speech were jointly decoded. The main issue with this purely top-down based approach is the search for binary mask labels is exponential. Barker et al. attempt to alleviate this problem by first grouping the T-F units and assigning labels to the overall groups. However, this ties performance to the accuracy of the initial groupings.

A more recent study [6] attempts to integrate the bottom-up spectral subtraction-based methods with a top-down approach. Using a conservative mask, a HMM lattice was generated containing speech hypotheses at each frame. The lattice is then rescored as the mask is updated based on comparisons between the noisy signal and the HMM models. While the system showed promising performance on a connected digit recognition task, the performance of our implementation degraded on larger vocabulary tasks. Our method is most similar to this system, but decouples the recognition and mask estimation and allows recognition to be performed in the cepstral domain with more robust features.

3. TOP-DOWN MASK ESTIMATION

We utilize a simple method for incorporating top-down linguistic information in the mask estimation process. In contrast to most methods that build a single general acoustic model, we build many simple acoustic models for specific sub-phonetic units. In order to perform masking in the spectral domain, we extract from each speech utterance a cochleagram (a popular signal representation widely used for IBM estimation and other purposes [7]). To generate the cochleagram, the signal is passed through a 64 channel gammatone filterbank to perform T-F decomposition. We also take the cube root of each T-F unit to limit the dynamic range.

We force-align our clean speech training data to obtain HMM state labels for each frame. For each state label, we take all training frames corresponding to that state label and create a mean spectral vector. Each mean vector is then unit-normalized as we are only interested in the general distribution of energy across frequency bands. Our mask estimation algorithm requires one further prior, a background energy prior. For our study, we utilize a simple vector that assumes equal energy in each frequency channel. However, a prior that incorporates knowledge about channel or noise characteristics could be used.

To illustrate the mask estimation process, we assume that for each frame of the input signal, we know the corresponding HMM state it is aligned with. To estimate the mask at each frame, we compare the state model associated with that frame to the background prior weighted by the relative energy at that frame. We define the masking criterion as

$$M(f,t) = \begin{cases} 1 & \alpha_{f,s_t} > \beta_f r_t^{\gamma} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$r_t = \frac{\text{average frame energy}}{\text{frame energy in frame } t} \quad (3)$$

where t is the time frame, f is the frequency band, α_{f,s_t} is the relative spectral energy in frequency f for the prior vector associated with HMM state s_t, β_f is the background prior, and γ is a factor used for nonlinearity.

The masking criterion seeks to keep any T-F units that should have strong energy based on the speech event that we are observing. This is accomplished by seeing if the energy percentage is greater for the speech prior than the weighted background prior. Weighting the background prior modifies the prior based on the amount of energy in a frame. We assume frames containing more energy are more likely to have strong speech energy also, though we recognize this assumption could be incorrect when the noise is temporally acute or impulsive. The γ value makes the weighting nonlinear so frames which deviate greatly from the mean are affected more. In this study a value of 2.5 is used, but different values have little effect on performance.

We want to emphasize that once the state model is chosen, the estimated mask is based solely on the expectation of which T-F units should have strong speech energy and the relative energy of the frame. No assumptions about noise are made and the individual T-F units of the data have no bearing on the mask estimation process. Once the mask has been estimated, we multiply it by the original signal and resynthesize the waveform. Standard ASR features can then be calculated from the enhanced waveform. The question of how to select the correct state model at each frame during test time still remains, which we will discuss in the next section.

4. EXPERIMENTAL SETUP

We use the HMM toolkit (HTK) [8] for our recognition system. The acoustic model consists of intra-word triphones; each triphone has three states, modeled by a mixture of 16 Guassians per state. A bigram language model is used during decoding. The CMU dictionary was used for our pronunciation dictionary. Our 39 dimensional feature vector is comprised of mean and variance normalized PLP features, including the delta and acceleration coefficients. While most systems using binary masking for speech enhancement use some type of missing feature recognition system [2, 9], we have found this can be unnecessary as long as the features are variance normalized [10].

All evaluations are performed on the Aurora4 corpus [11], a 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 at various SNR.

To illustrate the mask estimation process, we assumed we knew to which HMM state each frame of data corresponded. Obviously, in practice this will not be known *a priori*. Instead the correct state will have to be identified. While it seems

Mask Type	car	babble	restaurant	street	airport	train	avg
Baseline	27.3%	34.3%	36.7%	39.3%	35.0%	42.0%	35.8%
1-Best Estimate	25.2%	32.5%	35.5%	37.7%	33.4%	39.7%	34.0%
100-Best Estimate	23.9%	30.7%	34.3%	35.4%	33.8%	36.6%	32.5%
Oracle Models							
IBM	17.6%	15.8%	15.4%	19.5%	16.2%	19.6%	17.4 %
Clean Speech Oracle	19.0%	20.1%	24.1%	20.5%	22.6%	21.6%	21.3%
100-Best Oracle	20.5%	25.6%	28.1%	29.9%	27.3%	32.1%	27.3%

Table 1. Word error rates for various mask types on the Aurora4 dataset. Word Error Rate for clean speech is 9.8%

infeasible to select the correct HMM state from all possible choices, we can drastically reduce the number of competitors.

In this study we allow the baseline HMM to select a subset of possible states at each frame. To do this we generate an N-best list of size 100 using the unenhanced signal. Given the N-best list we can extract a list of possible states, or models, at each time frame. Using a list of 100 utterances produces an average of 1.7 model candidates per frame. For this pilot study, we use a simple method for choosing the state; this method is based on the assumption that high energy T-F units likely contain speech information and low energy units contain little speech information. We generate a noise estimate by averaging the first few frames of each utterance. Through spectral subtraction [12], we generate an estimate of the clean speech. While this approach would provide a poor mask estimation, we use it to identify strong and weak energy areas. Any T-F unit where the estimated SNR is greater than 10dB is assumed to be speech dominant and less than -10dB is assumed to be noise dominant. These units create a guide, sometimes referred to as a conservative mask, to compare candidate models against. We finally select the model whose mask most closely matches the guide. The hope is that if a model closely matches the labeling of the T-F units we are confident in, then the labelings of the unknown regions will be correct. Once a model is selected, we use its mask even where it differs from the conservative mask.

5. RESULTS

Table 1 presents word error rates when using various binary masks. The *Baseline* result uses no enhancement and provides a floor for our performance, while the *IBM* result provides a ceiling. First, we test the performance of our simple mask estimation method given perfect information. The clean speech utterance is force aligned to identify the ideal state model for each frame according to the ASR system. *Clean Speech Oracle* shows the performance of the estimated mask when using the ideal state. While performance is not as strong as the *IBM*, it produces a 40.5% error reduction over the *Baseline*.

The *Clean Speech Oracle* result confirms this approach to mask estimation can work. Next, we examine performance when the ideal state is unknown. Using the output of the baseline recognizer on noisy speech is the simplest method of selecting the candidate state at each frame. Since this only produces one candidate per frame, it does not require a method to select the correct model. While this approach, *1-Best Estimate*, should only reinforce the decisions made by the baseline recognizer, it does produce a small reduction in error.

Finally, we use the 100-best list to generate candidate states for each frame. Recall this requires our spectral subtraction based method to choose from the approximately 1.7 candidate states per frame. The *100-Best Estimate* produces a further reduction in error over the *1-Best Estimate* result for a total of 9% relative error reduction over the *Baseline*.

While the error reduction is significant over the baseline, a gap still exists with the *Clean Speech Oracle* result. A natural question is how much the simple model selection is impacting performance: the *100-Best Oracle* result replaces the model selection with the model closest to the *Clean Speech Oracle*. Performance is significantly better than our simple model selection method. While the list of candidate states could also be improved, the poor model selection would only be further exacerbated by increasing the number of possible models.

While our main goal is improved ASR performance, we also present results on a typical speech enhancement metric, SNR improvement. For comparative purposes, in Table 2 we list both the SNR improvements of our system and an MMSE-based system from Hendriks et al. [3], a current state of the art bottom-up system. Their system does not estimate a mask; rather it estimates the noise spectrum in order to extract the clean speech signal. The top-down approach performs favorably compared to the bottom-up approach; since one could envision using both techniques in a full system, the point here is not to claim one technique is better than the other, rather, this illustrates the power of top-down hypotheses, even ones with significant errors, in enhancing speech. While the goal of the Hendriks et al. system was not to improve ASR performance, we note that using their estimated clean speech in an ASR system did not improve performance over the *Baseline*.

6. CONCLUSIONS AND FUTURE WORK

We have outlined a simple ASR-driven top-down approach to binary mask estimation for ASR. Our approach utilizes hy-

Enhancement Technique	car	babble	restaurant	street	airport	train	avg
Hendriks et al. [3]	8.3	2.8	2.3	6.7	2.4	5.7	4.7
100-Best Estimate	10.9	3.1	2.3	7.1	2.8	6.0	5.4

Table 2. SNR improvements for several speech enhancement methods on the Aurora4 dataset.

potheses generated by an HMM recognizer to identify T-F units that are likely to be masked given the hypothesized subphonetic unit. The approach is agnostic to the underlying interference and is only concerned with the underlying speech.

Given perfect information, our approach can reduce word error by 40% over an unenhanced baseline. Using a 100-best list to generate hypotheses and a simple model selection method, we could reduce word error by 9% over an unenhanced baseline. While state of the art methods would outperform the result presented here, this pilot study demonstrates that top-down information can be useful and provides a simple framework for utilizing it.

Ultimately, we envision this approach to mask estimation to be an iterative process. The baseline recognizer will first generate hypotheses to be used for mask estimation. Then the recognizer will generate a list of improved hypotheses using the enhanced signal, similar to the idea presented in [13]. An iterative process would not only allow the mask estimation to improve recognition, but the recognition to improve mask estimation. While most research has focused on speech enhancement improving ASR performance, we have shown the reverse can be true; speech recognition, even when inaccurate, can improve mask estimation or speech enhancement.

The focus of future work will be on improving the model selection process, as it is the main bottleneck in the current system. While we claim the current approach is agnostic to the interference, we recognize that better model selection methods will likely utilize more information about the underlying signal. Current conventional mask estimation methods may even improve the model selection process.

Acknowledgements. This work was supported in part by NSF grant IIS-0643901 (CAREER).

7. 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] 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.

[3] R. C. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *Proceedings of IEEE ICASSP*, 2010, pp. 4266–4269.

[4] W. Kim and J. H. L. Hansen, "A novel mask estimation method employing posterior-based representative mean estimate for missing-feature speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1434–1443, July 2011.

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

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

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] D. P. W. Ellis, "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures," *Speech Communication*, vol. 27, pp. 281–298, 1999.