Improved Model Selection for the ASR-Driven Binary Mask

William Hartmann, Eric Fosler-Lussier

Department of Computer Science and Engineering
The Ohio State University, Columbus, OH, USA
hartmann.59@osu.edu, fosler@cse.ohio-state.edu

Abstract

In a previous study, we proposed an alternative masking criterion for binary mask estimation based on the underlying linguistic information. We estimated this mask by selecting from a set of candidate masks at each frame based on the hypotheses from an ASR system. Our previous system provided an 8% reduction in WER. In this work, we present an improved method for selecting the correct candidate mask at each frame, increasing the reduction in WER to 14%. Our new method uses a discriminative sequence model and provides a framework that can incorporate other mask estimations as features.

Index Terms: speech recognition, binary mask estimation

1. Introduction

One of the longstanding issues with automatic speech recognition (ASR) systems is the inclusion of additive noise in the input signal [1]. Many methods and systems have been proposed to compensate for this phenomenon; here we focus on speech separation techniques based on Computational Auditory Scene Analysis (CASA), in particular attempting to predict the Ideal Binary Mask (IBM) as a computational goal [2].

The IBM can be calculated from the spectro-temporal representation of a signal by computing the instantaneous SNR for each time-frequency (T-F) unit. Formally, we can define the IBM as the binary labeling $M(f,t)$ arising from the criterion:

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

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 1, which also corresponds to an SNR of 0 dB.

Given an IBM, the mask is then multiplied by the original signal in order to remove noise-dominant T-F units. The IBM represents a goal and requires *a priori* information in order to be calculated; in practice, the IBM must be estimated from the mixed signal. Many estimation methods have been proposed, but they typically only use low-level signal features [3]. Techniques which also incorporate speech information focus on general models of speech [4].

Very few methods have attempted to incorporate top-down linguistic information into the mask estimation process. Of those methods, they suffer from the inherent computational complexity of decoding over all possible mask patterns [5] and the use of spectral features for recognition [6]. In a previous study [7], we proposed an alternative masking criterion that forced the use of higher level linguistic information in the mask estimation process. Our estimation algorithm hinged on the ability to select the correct mask from a small number of competitors at the frame level. In this study, we propose an improved model selection method that significantly outperforms our previous published method.

In Section 2 we briefly describe the ASR-driven mask. Our improved model selection method is presented in Section 3. Section 4 contains a description of our experimental setup and results are reported in Section 5. Finally, conclusions and ideas regarding future work are discussed in Section 6.

2. The ASR-Driven Binary Mask

A typical HMM-based ASR system models speech using context-dependent phonetic units. These phonetic units can number in the thousands and provide very fine-grained linguistic information about the signal at the frame level. In our previous work [7], we used these phonetic units to define an alternative masking criterion. With an IBM, the mask at each T-F unit depends on the instantaneous SNR. The ASR-driven binary mask instead depends on the underlying phonetic unit for each frame and the distribution of energy in the input signal.

For each phonetic unit, we model the average distribution of energy seen in a training set. Our assumption is that T-F units which typically contain large amounts of speech energy for a phonetic unit should remain unmasked regardless of the energy in the noise source; masking T-F units that typically contain little speech energy should have little effect. We choose T-F units to

Triphone Context	Freq. Dependent	Feature Type
Observation Feature Functions		
Left Context	no	bias
Right Context	no	bias
Center Context	no	bias
Center Context	yes	bias
Center Context	yes	posterior
none	yes	bias
none	yes	posterior
Transition Feature Functions		
Center Context	no	bias
Center Context	yes	bias
none	yes	bias

Table 1: List of feature functions used in the sequence model. Triphone context refers to functions related to a portion of the triphone. Frequency dependence describes whether the function is general or has a version for each of the 64 frequency bins. Feature type refers to whether the value of the feature is a bias or an MLP posterior.

mask by comparing the model to a weighted background prior: the ASR-driven binary mask is defined as

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

$$r_t = \frac{\text{average frame energy}}{\text{frame energy in frame } t} \tag{3}$$

where t is the time frame, f is the frequency band, α_{f,p_t} is the relative spectral energy in frequency f for the phonetic unit p_t, β_f is the background prior (which assumes equal energy in every frequency band in this study), and γ is a factor used for nonlinearity. Based on results on a development set, we set $\gamma = 2.5$. The weighting r_t is used to control the aggressiveness of the masking based on the amount energy in a particular frame. A more detailed description of the ASR-driven binary mask can be seen in our previous study [7].

As with the IBM, we have described a masking criterion that relies on oracle information. In practice, a set of candidate phonetic units can be hypothesized for each frame via a lattice or n-best list from an initial ASR pass. Once the candidate phonetic units have been hypothesized, we select a single candidate at each frame; this process is described in the following section.

3. Candidate Model Selection

We have a set of candidate models and masks for every frame. Since the mask is defined in terms of the frame energy and the phonetic unit, any candidate model corresponds to only one possible mask. The task is to select one candidate model at each frame in order to estimate the complete mask. Our previous method selected the candidate mask that most closely matched a simple spectral subtraction (SS) based mask. This method has several obvious issues: a SS-based mask estimation will contain many errors compared to the IBM; also, since our masking criterion is inherently different than the IBM, basing our estimate off of a method attempting to estimate the IBM may produce poor results.

A measure is needed to directly estimate the closeness of a candidate model to the true ASR-driven mask at a given frame: we train a simple multi-layer perceptron (MLP) to produce a posterior probability of a given T-F unit being masked. The three features used by our MLP are the SS-based instantaneous SNR estimate, the weighted background prior, and the average value for all candidate masks at that T-F unit. Since the MLP is trained with the ASR-driven mask as its objective, it provides a much better estimate than the SS-based mask.

However, the MLP treats each T-F unit decision as an independent problem. We propose a method for model selection that uses the inherent sequence information provided by the hypothesized phonetic units; a discriminative linear chain sequence model is trained using the structured perceptron algorithm [8]. Our method differs slightly from typical structured prediction problems due to the size of our label space.

The possible labels for any frame are the cross product between the number of phonetic models and the possible mask patterns at each frame. In our study, we have approximately 20,000 phonetic states and 2^{64} possible masks. In a typical sequence modeling task, each label or pair of labels has an associated set of feature functions. This will not work in our domain due to both computational and data sparseness reasons. Instead, our feature functions are associated with more general information about the labels. For instance, we have feature functions which will be nonzero when the candidate mask contains a 1 in the ith frequency band. A full listing of our feature functions can be seen in Table 1. Feature functions come in two varieties: bias terms are always set to one, and the weight models the feature prior. MLP posteriors represent the second type, and give an estimated masking strength at the corresponding frequency channel.

While any feature function will be nonzero for a large number of labels, a unique set of feature functions will be nonzero for every label. We are able to use this type of model because we do not need to decode over the entire set of labels. The first pass through the ASR system provides a small number of candidate models at each frame. The sequence model allows the model selection process to make use of a rich set of information regarding the correlation between mask values and phonetic units.

The estimated mask is the mask sequence that maximizes

$$\underset{y}{\operatorname{argmax}} \sum_i \sum_k \alpha_k f_k(y_i, y_{i-1}, x_i) \tag{4}$$

where i is the frame index, k is the feature function index and α_k is the weight associated with the feature function $f_k(\cdot)$. The x_i vector contains the posteriors predicted by the MLP at that frame. The y vector represents the label sequence where each frame is given a phonetic model label and a frame-level mask. Using the viterbi algorithm, the sequence of candidate models which maximizes the score can quickly be found.

Notice that there is a slight disconnect between our label space and our objective. In the end, we only care about obtaining the correct mask, the associated phonetic units are irrelevant. We recognize this issue and make a small change to the perceptron update rule. When the hypothesized mask matches the true mask at a particular frame, we do not update any of the associated feature functions regardless of the hypothesized phonetic unit. In this way, we do not penalize any of the weights as long as the predicted mask is correct.

4. Experimental Setup

We report results on the Aurora4 corpus [9], a 5000 word closed vocabulary task. The corpus is a modified version of the Wall Street Journal (WSJ0) database with a variety of noise types added to clean speech recordings. All HMMs used in our experiments were created using HTK [10]. Acoustic models trained on clean speech consist of tied-state, intra-word triphones with three states per triphone. Each state is modeled by a mixture of 16 Gaussians. A standard bigram language model was used for decoding and the CMU dictionary was used for pronunciations. Mean and variance normalized PLP features, including delta and acceleration coefficients, comprised the 39-dimensional feature vector. Most ASR systems using binary masks incorporate a missing feature recognition methodology [3, 5, 6], but we have previously found this to be unnecessary if the acoustic features are variance normalized [11].

The trained acoustic models were used not only for the final recognition results, but also for data alignment and computation of word lattices. Each training sentence was force-aligned to provide a state label for each frame. The phonetic prior models described in Section 2 were created by calculating the mean energy distribution in the cochleagram [12] domain of all the data for each state label.

We also used the training data to train both the MLPs and the discriminative sequence model. A lattice was generated for each training utterance, providing a set of hypothesized phonetic units at each frame. The training lattices were trained from clean speech since generating multiple lattices for various noise conditions would be expensive. To create the SS-based noise estimate and weighted background prior features used by the MLP, each training utterance was mixed with a noise source form the Noisex92 database [13] at a random SNR between 20 and 0 dB SNR. A total of 64 MLPs, corresponding to the 64 frequency channels used to calculate the cochleagram, were trainied using the softmax criterion and consisted of 3 input features, 9 hidden units, and 2 posterior outputs. The discriminative sequence model, trained to maximize mask accuracy, also used the same training data with the MLP posteriors as features. Only one iteration of training was used as further iterations did not improve performance on a development set.

During testing, lattices were generated from the unenhanced speech to generate phonetic unit hypotheses for each frame. MLP posteriors were generated using the data and the mask was finally decoded using the trained sequence model. The mask was multiplied by the original data in the cochleagram domain and resynthesized back into the time domain. Acoustic features were then calculated for recognition.

5. Results

Word error rate (WER) results are reported in Table 2 on the Aurora4 dataset. *Baseline* refers to recognition on the unenhanced noisy speech. *Oracle ASR-Driven Mask* demonstrates a performance ceiling for our system by selecting the candidate mask at each frame which most closely matches the true model found by force-aligning the clean speech. The *SS-based Estimate* is our previously published system and the sequence model *(SM)-based Estimate* is the system using the improved model selection presented in this study. Perfect model selection results in an average WER reduction of 31%; the improved model selection increases WER reduction to 14% compared to 8% in our previous study.

We note that results differ slightly (but insignificantly) from the previous study as recognition parameters such as the language model weight and phone insertion penalty were tuned on the development system. Also, in the previous study, we used a 100-best list instead of a lattice to produce candidates. With the improved model selection, we are now able to make use of the extra model hypotheses the lattice provides.

The new system improves upon the previous system in several ways. Using the MLP to generate an estimated mask gives a better first estimate than the SS-based mask, likely because the MLP was trained to predict the ASR-driven binary mask while the SS-based mask more closely approximates the IBM. The sequence model is also able to make use of transition information. The models we use implicitly contain information about transitions, so it is not surprising incorporating this information improves performance.

Our previous system worked by selecting the model that most closely matched a SS-based mask estimation. It treated each frequency channel as equally important. It is likely that the relative importance of different frequency channels will vary with phonetic context; the SM-based

Mask Type	car	babble	restaurant	street	airport	train	avg	rel. imp.
Baseline	27.7%	34.7%	37.3%	39.9%	35.5%	42.2%	36.0%	–
SS-Based Estimate	24.4%	30.8%	35.7%	35.4%	34.4%	37.2%	33.0%	+8.3%
Direct MLP Estimate	25.1%	30.2%	35.7%	35.2%	32.9%	36.4%	32.6%	+9.4%
SM-Based Estimate	24.0%	28.6%	32.6%	34.6%	31.3%	34.3%	30.9%	+14.2%
Oracle Models								
Oracle ASR-Driven Mask	*20.7%*	*24.2%*	*27.3%*	*24.9%*	*26.1%*	*26.5%*	*24.9%*	*+30.8%*
Ideal Binary Mask	*17.6%*	*16.8%*	*15.2%*	*18.1%*	*15.6%*	*19.1%*	*17.1%*	*+52.5%*

Table 2: Word error rates for various mask types on the Aurora4 dataset. The final column shows the relative improvement of each mask compared to the baseline in terms of average word error rate. Word Error Rate for clean speech is 9.8%

system can incorporate this idea via the feature functions associated with each individual frequency band.

The *Direct MLP Estimate* result shows that the MLP estimate is a better mask estimate than the our previous SS-based mask for the purposes of ASR. However, incorporating the MLP-based estimate into our new system provided a significant performance improvement. We expect that when confronted with a stronger baseline subsequently, we could incorporate it as a feature in our framework and provide further performance improvements.

6. Conclusions

We have presented an improved method for model selection for the estimation of the ASR-driven binary mask. The method incorporates an MLP-based estimation into a discriminative sequence model. Our new method takes advantage of the available transition information and the relative importance of different frequency channels for different phonetic contexts. The framework we have proposed should also be able to incorporate information from stronger baseline systems as features.

Our future work will take two directions. The estimation process is only as good as the initial set of hypotheses. We believe an iterative process may further improve performance. Once the first mask has been generated, a new set of hypotheses can be generated from the enhanced speech. A new mask can now be estimated from the improved hypotheses. This idea is similar to the one proposed in [14].

Our framework relies on associating a frame-level binary mask to each phonetic unit. A simple method based on the expected distribution of energy for a given phonetic unit results in a mask that performs well, but not as well as the more traditional IBM (compare the IBM and ASR-driven oracle masks in Table 2). In future work, we will investigate better ways of associating binary masks to phonetic units that more closely match the IBM.

7. Acknowledgements

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

8. References

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

[2] 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.

[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] 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] W. Hartmann and E. Fosler-Lussier, "Asr-driven top-down binary mask estimation using spectral priors," in *Proceedings of IEEE ICASSP*, pp. 4685–4688, 2012.

[8] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in *Proceedings of EMNLP*, 2002.

[9] 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.

[10] 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

[11] W. Hartmann and E. Fosler-Lussier, "Investigations into the incorporation of the ideal binary mask in asr," in *Proceedings of IEEE ICASSP*, pp. 4804–4807, 2011.

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

[13] 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.

[14] 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.