

STRUCTURE IN TIME-FREQUENCY BINARY MASKING

A Dissertation
Presented to
The Academic Faculty

by

Abigail A. Kressner

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2015

Copyright © 2015 by Abigail A. Kressner

STRUCTURE IN TIME-FREQUENCY BINARY MASKING

Approved by:

Christopher J. Rozell, Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Jörg M. Buchholz
Department of Linguistics
Macquarie University

David V. Anderson
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Pamela T. Bhatti
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Mark A. Clements
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Date Approved: 12 March 2015

ACKNOWLEDGMENTS

I wish to thank my committee members who have each guided me and supported me in their own way during my studies. Chris Rozell has not only taught me how to conduct research, but he has also showed me how one can lead a team of very diverse individuals in an effective way. Chris has been a champion for me from start to finish during this very long endeavor, and I am eternally grateful. Jörg Buchholz graciously welcomed me with open arms to Sydney to embark on an unforgettable journey, both personally and professionally. Without doubt, the last nine months at the National Acoustic Laboratories has been one of the most rewarding research experiences one could ask for. Mark Clements and David Anderson have given me nothing but relentless support and encouragement since before even starting at Georgia Tech. Although my face time with both Mark and David was limited, I felt their presence always and knew I could call on them if I needed to. Finally, although almost certainly without realizing it, Pamela Bhatti provided for me an essential strategy for standing strong during moments of internal weakness.

I am also eternally grateful to each and every one of my dearest friends and family. I will forever be thankful to my parents because I would not be here, writing this, living the life I live, and doing the things I do, if it weren't for them. I would also like to thank my sisters, Lyndsay and Tamie, for bothering to teach me silly things like long division way back when.

Last but certainly not least, I owe my deepest gratitude to both my twin sister Josie and my partner Adam for their unrelenting love and support. This journey has been absolutely unforgettable with the two of them by my side, picking me up from the lows and cheering me on during the highs. It has been a remarkable thing to have both of their understanding and encouragement surrounding me every single day.

This work was supported in part by the National Science Foundation's Graduate Research Fellowship (GRFP) and Graduate Research Opportunities Worldwide (GROW) programs, as well as the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	xi
SUMMARY	xiii
I INTRODUCTION	1
1.1 Objectives	3
1.2 Outline	3
II SPARSITY-BASED BINARY MASK ESTIMATION	5
2.1 Introduction	6
2.2 Background	6
2.3 Algorithms	8
2.4 Methods	11
2.5 Results	12
2.6 Discussion	16
III SPARSITY-BASED SPEECH CODING STRATEGY	19
3.1 Introduction	20
3.2 Algorithms	20
3.3 Methods	25
3.4 Results	27
3.5 Discussion	28
IV STRUCTURE IN BINARY MASK ESTIMATION ERROR	31
4.1 Introduction	32
4.2 Description of the model	32
4.3 Methods	35

4.4	Results	41
4.5	Discussion	45
V	COCHLEAR IMPLANTS AND BINARY MASK ERRORS . .	51
5.1	Introduction	52
5.2	Methods	52
5.3	Results	59
5.4	Discussion	61
VI	OBJECTIVE ANALYSIS OF STRUCTURE IN MASKS	65
6.1	Introduction	66
6.2	Methods	67
6.3	Results	73
6.4	Discussion	79
VII	CONCLUSION	85
7.1	Limitations	87
7.2	Future directions	87
	REFERENCES	89
	VITA	95

LIST OF TABLES

1	Error rates and STOI predictions	18
2	Biographical data for the CI recipients	56
3	List of estimation algorithms and their primary differences.	68
4	List of model parameters.	72

LIST OF FIGURES

1	Time-frequency representations, masks, and spikegrams for example sentence	10
2	H-FA analysis of sparsity-based estimation algorithms	12
3	STOI predictions for sparsity-based estimation algorithms	14
4	Error rate analysis for sparsity-based estimation algorithms	15
5	Example masks with highlighted false positive and negative errors .	17
6	Spectrograms and electrodograms for an examle sentence	24
7	Analysis of MP weights and number of active electrodes	25
8	Objective analysis of vocoded speech	27
9	Results from listener study with vocoded speech	28
10	Illustration of a graphical model for binary masks	33
11	Example masks comparing error structure	34
12	Model-generated masks with varying levels of false positives and clustering	39
13	Model-generated masks with varying levels of false negatives and clustering	40
14	Model-generated masks with varying levels of false positives, false negatives, and clustering	40
15	Mean word recognition scores demonstrating the effect of structure in false positive errors	42
16	Mean word recognition scores demonstrating the effect of structure in false negative errors	43
17	Mean word recognition scores demonstrating the effect of interacting errors and structure	44
18	Contour plot of word recognition for unstructured and structured errors with normal-hearing listeners	46
19	Example unstructured and structured binary masks with the respective CI spectrograms and electrodograms	53
20	Speech recognition scores for binary-masked speech with CI recipients	60

21	Contour plot of word recognition for unstructured and structured errors with CI recipients	62
22	Example masks demonstrating cluster orientation	71
23	Distribution of error rates among estimated binary masks	74
24	Distribution of clustering across estimation algorithm	75
25	Distribution of decoupled clustering across estimation algorithm . .	76
26	Comparison of contour plots based on behavioral data and objective measures	78
27	Histograms of local SNR	81

SUMMARY

Understanding speech in the presence of interfering background noise is a challenge for normal-hearing and impaired-hearing listeners alike. However, it has been shown that intelligibility outcomes can be improved in these situations with both noise reduction and cochlear implant channel selection strategies based on the ideal binary mask processing approach. This approach uses prior information about the time-frequency composition of the target and interfering signals to retain only the temporal and spectral regions of the mixture signal that contain more energy from the target signal than the interferer signal. In realistic scenarios where prior information is unavailable however, the ideal binary mask must instead be estimated.

To date, researchers have primarily used uniformly random, uncorrelated mask errors and independently presented error types (i.e., retaining interferer-dominated units that should have been removed or removing target-dominated units that should have been retained) to characterize the influence of estimation errors on intelligibility. However, practical binary mask estimation algorithms produce masks that contain errors of both types and with nontrivial amounts of structure. The main contributions of this thesis are to introduce a novel investigation framework for binary masks and to present listener studies with both normal-hearing listeners and cochlear implant recipients that use this framework to illustrate how interactions between error types and structure affect speech recognition outcomes.

The results specifically demonstrate that clustering of mask errors reduces speech recognition. Moreover, while previous research suggested that incorrectly retaining interferer-dominated units was more detrimental to intelligibility than removing target-dominated units, the results from these listener studies indicate that incorrectly removing target-dominated units can be equally detrimental. The studies in this thesis jointly establish that structure substantially impacts how much classification error is acceptable, and since realistic noise reduction and channel selection strategies will make errors in structured ways, it is important to consider the structure when analyzing potential outcomes.

CHAPTER I

INTRODUCTION

Listening to speech in the presence of background noise is a challenging task for normal-hearing listeners, and it is even more challenging for impaired-hearing listeners and cochlear implant (CI) recipients. Furthermore, speech communication in realistic environments (e.g., multiple sound sources in a reverberant room) is often problematic for impaired-hearing listeners and CI recipients even after compensating for their decreased sensitivity to sound. Researchers have attempted to improve speech understanding in these environments in a variety of ways, but traditional approaches on single-channel signals have had limited success in actually improving the intelligibility of speech (Hu and Loizou, 2007, Loizou and Kim, 2011). Because of this lack of success, finding ways to improve the intelligibility of noisy speech remains an active area of research.

The development of the ideal binary mask (IBM) algorithm provides evidence that with prior knowledge of the target speech and interfering noise signals, even very basic processing can lead to the improvement of intelligibility for normal-hearing listeners (Roman et al., 2003, Wang, 2005, Brungart et al., 2006), and to an even further extent, for impaired-hearing listeners (Anzalone et al., 2006) and CI recipients (Hu and Loizou, 2008, Mauger et al., 2012). The general approach of the IBM is to first create a binary mask, which is defined as a matrix of binary gain values. Then, using separate time-frequency (T-F) representations of the target and interferer signals, the local signal-to-noise ratio (SNR) is computed within each T-F unit. Mask units that correspond to T-F units with a local SNR above a pre-defined threshold are assigned a value of one and zero otherwise. This mask is subsequently applied to the mixture before recombination with a synthesis filterbank. Alternatively, in the framework of CIs the mask is either applied to the mixture before channel selection or it is implemented as the selection criteria for channel selection itself.

Even though the IBM requires unrealistic prior knowledge of the target and interferer signals, the significant intelligibility increase makes it a desirable benchmark for speech de-noising and channel selection algorithms. Much work has gone into estimating the IBM when prior knowledge is unavailable. By formulating the mask estimation problem as a classification problem and using machine learning to form the masks, Kim et al. (2009) were the first to show with listener studies that estimated masks can improve speech intelligibility in noise. Subsequent to this, many researchers have proposed new approaches using more recent machine learning techniques (e.g., Kim and Loizou, 2010a, May and Dau, 2013, Han and Wang, 2012, Wang and Wang, 2012, 2013). Recently, Healy et al. (2013) showed with

listener studies that a method based on a state-of-the-art machine learning techniques can improve speech intelligibility specifically in impaired-hearing listeners. However, May and Dau (2014b) have subsequently shown that introducing variations in the noise unseen by a classifier during training (even different realizations of the same noise) can yield dramatically reduced mask estimation performance and, therefore, that the intelligibility improvements demonstrated thus far with these estimation algorithms may be valid in only limited conditions.

While developing these binary mask estimation algorithms, it is desirable to evaluate performance without having to run listener studies. Knowing basic principals of how different types of errors influence the intelligibility outcomes would be a valuable guide in the development process. In an effort towards this, Li and Loizou (2008) looked at some of the factors that influence the intelligibility of binary-masked speech. They found that intelligibility was greater than 90% when overall error rates for uniformly random errors was at or below 10%. Furthermore, they found that when only uniformly random false positives (interferer-dominated units incorrectly labeled as target-dominated) are present, false positive error rates as low as 20% are detrimental to intelligibility. Contrastingly, when only uniformly random false negatives (target-dominated units incorrectly labeled as interferer-dominated) are present, false negative error rates as high as 60% still yield high intelligibility. Based on the assumption that false positives and false negatives act independently, Li and Loizou (2008) argue that false positives are substantially more detrimental to intelligibility than false negatives.

With this knowledge in hand, it is a logical next step to try to develop novel estimation algorithms that can more effectively and consistently improve speech intelligibility. Since the IBM creates a T-F representation that is more sparse (i.e., has fewer non-zeros) than the original mixture signal by preserving only T-F regions that are target-dominated, sparse coding models make for a promising estimation strategy candidate. Moreover, given the recent advances in the signal processing community revolving around this notion of sparsity and the fact that the signal processing community is actively developing methods to compute sparse approximations efficiently (i.e., in real-time and using low power), there is a possibility that sparsity-based approaches may be utilized for low-power and time-efficient estimates of the IBM.

However, the research to date lacks enough understanding about the factors that influence the outcomes of time-frequency gain manipulation to properly guide the development of a sparsity-based binary masking algorithm. In particular, the current set of guidelines are based on research that ignored two basic characteristics that are almost certainly true of all practical binary mask estimation algorithms. First, estimated masks contain both types of errors (false positives and false negatives) simultaneously. Second, mask errors are not uniformly distributed in time and frequency, and instead are structured in a variety of ways. Despite these characteristics being prominent features of practical binary mask estimation algorithms,

it is uncertain how interactions between error types and structure in mask errors affects intelligibility performance for binary masking approaches in the context of both noise reduction and cochlear implant channel selection.

1.1 Objectives

The main objective of this thesis is to demonstrate the influence of structured and unstructured errors on speech intelligibility in T-F binary masking. In the path towards defining this research direction, contributions were also made to the development of a binary mask estimation algorithm and a channel selection strategy that both exploit the sparse coding model. The thesis focuses on the following specific aims:

1. Develop noise reduction and channel selection algorithms based on binary masking that leverage the sparse coding model.
2. Develop an investigation framework for characterizing the effects of binary mask errors.
3. Demonstrate the influence of structure in binary mask errors on speech intelligibility.

1.2 Outline

This thesis consists of the main contributions in Chapters 2 through 6 and a concluding discussion in Chapter 7 which summarizes the main findings of the thesis and discusses the collective impact of the work from a broader perspective. Each of the specific aims are addressed as follows:

Develop binary mask algorithms that leverage sparsity: This aim is addressed in Chapters 2 and 3. In Chapter 2, a new binary mask estimation approach that uses a sparse approximation algorithm called matching pursuit (MP) is proposed and evaluated with prevalent binary mask objective measures. Subsequently in Chapter 3, a channel selection algorithm that also uses MP is evaluated first objectively with vocoded speech and an objective measure and then subjectively with vocoded speech and normal-hearing listeners. Despite being able to satisfy the criteria implied by Li and Loizou (2008) for effective binary mask estimation, the speech intelligibility outcomes of the proposed algorithms are collectively lower than expected. As a result, the need for more specific and reliable design criteria is discussed.

Develop investigation framework for binary masks: This aim is primarily addressed in Chapters 4 and 6. While the literature to date largely addresses straightforward forms of classification accuracy (e.g., Anzalone et al., 2006, Li and Loizou, 2008), the effect of the distribution of different types of mask errors on speech intelligibility remains unknown. Understanding how specific kinds of structure influence

speech intelligibility outcomes will aid in the future design of binary-masking algorithms. To facilitate investigations of structure in mask errors, a probabilistic model is developed that captures the conditional dependencies within a binary mask and between the binary mask and its corresponding ideal version. This model is first introduced in Chapter 4 and then extended in Chapter 6.

Demonstrate influence of structure in binary mask errors: This aim is addressed collectively in Chapters 4 through 6. Each of these chapters use the investigation framework for binary masks to either artificially generate masks with specific error characteristics or to learn the error characteristics of authentically estimated binary masks. With artificially generated masks, Chapter 4 demonstrates the effect of structure on the ability of normal-hearing listeners to tolerate binary mask errors. Following this, Chapter 5 also uses artificially generated masks to establish the effect of both structured and unstructured errors on speech recognition in CI recipients. Finally, Chapter 6 uses the model to more thoroughly evaluate structure in authentic binary masks from a collection of existing estimation algorithms in the literature to assess the implications of the listener studies in Chapters 4 and 5 on real-world algorithm development based on the binary masking approach.

CHAPTER II

A BINARY MASK ESTIMATION ALGORITHM BASED ON SPARSITY

The main objective of this chapter is to introduce a novel binary mask estimator based on a simple sparse approximation algorithm. The results suggest that sparsity-based approaches can do reasonably well, at least in terms of classification accuracy, when the sparse coding model appropriately captures the statistical differences between the target and interferer. Since there is already progress towards solving sparse approximations with real-time, low-power architectures, sparsity-based approaches are worthy of further consideration. Looking forward though, it will be necessary to first define more specific design criteria for estimated binary masks in order to develop more effective estimation algorithms.

Parts of this chapter are based on the papers that were published as “A novel binary mask estimator based on sparse approximation” by Abigail A. Kressner, David V. Anderson, and Christopher J. Rozell in the Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013 and as “Causal binary mask estimation for speech enhancement using sparsity constraints” by Abigail A. Kressner, David V. Anderson, and Christopher J. Rozell in the *Proceedings of Meetings on Acoustics*, Montreal, Canada, June 2013.

2.1 *Introduction*

Wang (2008) published a detailed review of T-F masking for speech separation and discussed at length its origin and its potential for application in hearing aids. Since then however, a lot of progress has been made in understanding when and why the IBM improves speech intelligibility (e.g., Li and Loizou, 2008, Hu and Loizou, 2008, Wang et al., 2009, Kim et al., 2009, Kjems et al., 2009, Loizou and Kim, 2011, Boldt, 2011, Roman and Woodruff, 2011, Healy et al., 2013) and how the IBM relates to perceptual effects such as quality and listening effort (e.g., Brons et al., 2012, Mauger et al., 2012, Ahmadi et al., 2013).

Drawing on the success of Kim et al.’s (2009) estimation algorithm that employed a Gaussian mixture model (GMM) to classify T-F units as target- or interferer-dominated, Hu and Loizou (2010) and May and Dau (2013) proposed amendments to the approach in order to increase the algorithm’s generalization to novel signals. Furthermore, May and Dau (2013) proposed a novel method to exploit T-F context by considering adjacent T-F units during classification instead of considering each unit independently. Both modifications have demonstrated improvement to classification accuracy. GMMs are generative models, which may not necessarily be the best choice for T-F unit classification. In contrast, state-of-the-art systems in the machine learning and automatic speech recognition community typically utilize discriminative models for classification tasks. In the context of IBM estimation, recent studies suggest that the GMM models can be replaced by discriminative learning methods such as support vector machines (SVMs) (Han and Wang, 2012), conditional random fields (CRFs) (Wang and Wang, 2012), and deep neural networks (DNNs) (Wang and Wang, 2013).

However, in order to achieve higher classification accuracy, these estimation approaches based on machine learning techniques often train classifiers independently for male and female utterances, or they use entirely different classification techniques for voiced and unvoiced segments. Furthermore, many of the algorithms require independent classifiers for each frequency sub-band or significant, non-causal post-processing. Each of these properties can make implementation for applications such as hearing aids and cochlear implants particularly challenging, since they require real-time and low-power processing. The main contribution of this work is to present a sparsity-based binary mask estimation algorithm that is amenable to real-time, low-power implementation.

2.2 *Background*

Many recent advances in signal processing for both signal modeling and signal acquisition have revolved around sparse coding (Baraniuk et al., 2010). Sparse coding models treat a signal as a linear combination of elements from a dictionary, and sparse approximation uses these models to find approximations to signals using as few of the dictionary elements as possible. The convolutional form of the sparse

coding model is given as

$$x(t) = \sum_{m=1}^M \sum_{i=1}^{n_m} s_{m,i} \phi_m(t - \tau_{m,i}), \quad (1)$$

where an audio signal, $x(t)$, is represented by a linear superposition of a basic set of dictionary elements, $\phi_1(t), \dots, \phi_M(t)$, which can be positioned arbitrarily and independently in time, and $\tau_{m,i}$ and $s_{m,i}$ are the temporal position and amplitude of the i^{th} instance of the kernel $\phi_m(t)$, respectively. The notation n_m indicates the number of instances of $\phi_m(t)$, which need not be the same across kernel functions. To code signals efficiently, one generally needs to find the optimal set of $\tau_{m,i}$ and $s_{m,i}$ (known as *encoding* or *inference*), as well as the optimal set of $\phi_m(t)$ (known as *learning*). Note that inference is done on a signal-by-signal basis, whereas learning is done once on a collection of signals to capture signal statistics. To build intuition, the signals that are most interesting in the real world generally have low-dimensional underlying structure, and this approach seeks to efficiently represent that essential information in just a few coefficients (i.e. the computational objective for encoding is to minimize the error of the representation while maximizing coding efficiency). For context, this is the model underlying the recent well-publicized signal processing results in “compressed sensing” (Candes and Wakin, 2008).

Alongside these advances in signal modeling, there is mounting evidence from computational neuroscience that neural systems may also use sparse coding to represent sensory information. For example, when sample training data of natural images and speech signals are used to learn optimal filters for the sparse coding model above (also called *efficient filters*), the results quantitatively match the response properties of visual receptive fields (i.e., gabor-like) (Olshausen and Field, 1996) and auditory nerve fibers (i.e., gammatone-like) (Smith and Lewicki, 2006), respectively. These filters are matched to the statistics of the signals of interest and are efficient at capturing the salient structure of the respective signal types. In fact, these ideas extend beyond physiology to perception. A recent study shows that distorted speech (noise-vocoded with only six filters) is more intelligible to listeners when the vocoder uses efficient filterbanks versus other common constructions (e.g., FFT/linear, cochleotopic/logarithmic) (Ming and Holt, 2009).

A plethora of algorithms exist for inferring the optimal set of coefficients ($\tau_{m,i}$ and $s_{m,i}$ in Eq. 1). Generally, algorithms must balance a trade-off between computational complexity and optimality. One very basic approach to efficient audio coding is filter and threshold (FT) (Smith and Lewicki, 2005); FT is a causal approach that achieves efficiency primarily by making use of a filterbank based on the human cochlea. It chooses coefficients based on the values and positions of the filter response magnitudes that exceed a preset threshold. This method reduces the information rate in comparison to the full convolutional representation, but since it does not take the correlations between basis functions into account, large amounts of redundancy remain. MP, on the other hand, is a computationally simple, greedy algorithm

designed to take these correlations between basis functions into account (Mallat and Zhang, 1993). For the convolutional model in Eq. 1, MP first chooses the time-shifted basis that has the largest inner product with the signal, then subtracts the contribution due to that basis, and repeats the process iteratively until the signal is satisfactorily approximated.

Another class of sparse coefficient recovery algorithms that are of interest are the locally competitive algorithms (LCAs) (Rozell et al., 2008). LCA is a neurally plausible algorithm that is based on the principles of local competition and thresholding; each neuron in a population of neurons models a basis function in an over-complete dictionary, and these neurons continually compete with each other to represent the input signal. This population of neurons is described by the dynamics of a set of non-linear ordinary differential equations (ODEs) that govern the internal state (membrane potential) and external communication (short-term firing rate) of each of the neurons in the population. The system of ODEs employ only simple computations that correspond to basic analog elements (e.g. resistors, capacitors, and amplifiers), which makes them amenable to analog implementation (Rozell et al., 2008). The fact that LCAs are amenable to analog architectures makes them a promising candidate for implementing in real-world applications many of the modern signal processing techniques that rely on sparse approximation. Toward this effort, Shapero et al. (2012) implemented a small-scale LCA on a low-power Field Programmable Analog Array (FPAA), and with this FPAA, they were able to reproduce with reasonable accuracy the outputs of a digital solver to a sample optimization program. Based on the trends of the performance, they predict that large-scale analog LCA systems (i.e. systems scaled to sizes that are appropriate for non-trivial problems) will compute solutions that compare favorably to digital solvers, but in about one-tenth of the time while using a small fraction of the power (Shapero et al., 2012, 2013). Moreover, Balavoine et al. (2012) showed that the convergence of the LCA is stable, that it is also globally convergent when the objective function is unique, and under some restrictions, that it converges exponentially fast with an analytically bounded convergence rate. Furthermore, Charles et al. (2012) demonstrated that LCAs can solve a wide-variety of sparsity-based probabilistic inference problems (e.g. ℓ_p norms with $0 \leq p \leq 2$, modified ℓ_p -norms, block- ℓ_1 norms, and re-weighted algorithms) that appear in the signal processing and statistics literatures.

2.3 Algorithms

Given the collection of progress towards low-power and real-time analog implementations of sparse approximation, sparsity-based solutions to existing problems in speech de-noising and coding have significant potential given that they are well-suited to overcome some of the barriers that currently exist in hearing aids and CIs. In this work, focus is placed on employing the FT and MP algorithms instead of LCA however since these algorithms are more commonly known.

2.3.1 Matching pursuit binary mask

To compute the proposed matching pursuit binary mask (MPBM), a sparse approximation is obtained for an arbitrary mixture signal using MP and a dictionary consisting of gammatones. Since speech is efficiently encoded with a gammatone dictionary, initial iterations in MP will likely choose coefficients that approximate speech energy rather than noise. Therefore, by choosing a suitable stopping criteria for MP, the reconstructed signal will largely contain gammatones that fall in the T-F regions of the target speech. Instead of actually synthesizing the approximated signal however, the sparse coefficients from MP are utilized to identify the target-dominated T-F regions in the following way.

Each of the sparse coefficients correspond to the placement of a gammatone in the reconstructed signal at a specific time-shift, and since gammatones are localized to small windows of time and frequency, they contain energy in a very specific localized T-F region. Therefore, the MPBM method assumes that the T-F units that contain significant energy for each of the chosen gammatones also contain the target. Thus, the T-F units in the binary mask that correspond to the T-F units that contain significant energy from the MP coefficients are given a value of one.

To illustrate this approach, Fig. 1a shows an example noisy sentence in gray and the clean sentence superimposed in black. Fig. 1b shows a “gammatonegram” (i.e., a spectrogram constructed with a gammatone filterbank) of the noisy signal, and Fig. 1c illustrates the IBM using a 0 dB threshold.

Fig. 1d shows the MP “spikegram” (each circle represents a coefficient in Eq. 1 and the size of the circle is proportional to the amplitude) for the example sentence and Fig. 1e shows the MPBM. To further illustrate the connection between MP coefficients and the resulting mask, a single MP coefficient is isolated in the spikegram in Fig. 1d in dark gray and the corresponding gammatone is superimposed onto the noisy and clean waveforms in Fig. 1a. Finally, the T-F region that corresponds to the single coefficient is shown in Fig. 1e in dark gray. Since this estimation process is clearly non-causal, a key factor going forward is to alter this approach to use more realistic frame-based, causal computations.

2.3.2 Frame-based matching pursuit binary mask

The general concept of the frame-based matching pursuit binary mask (fMPBM) remains the same as MPBM in that the sparse coefficients from MP are utilized to identify the target-dominated T-F regions. However, instead of obtaining a sparse approximation of the entire signal at once, approximations for short, overlapping segments (i.e., frames) of the mixture are obtained. Then, for each frame, the binary mask for all correlated channels is set to one for the duration of the frame. The frame-based matching pursuit (fMP) spikegram for the example sentence is shown in Fig. 1f and the resulting fMPBM in Fig. 1g. Even though MP is running on short segments of the speech rather than the entire signal at once, the resulting

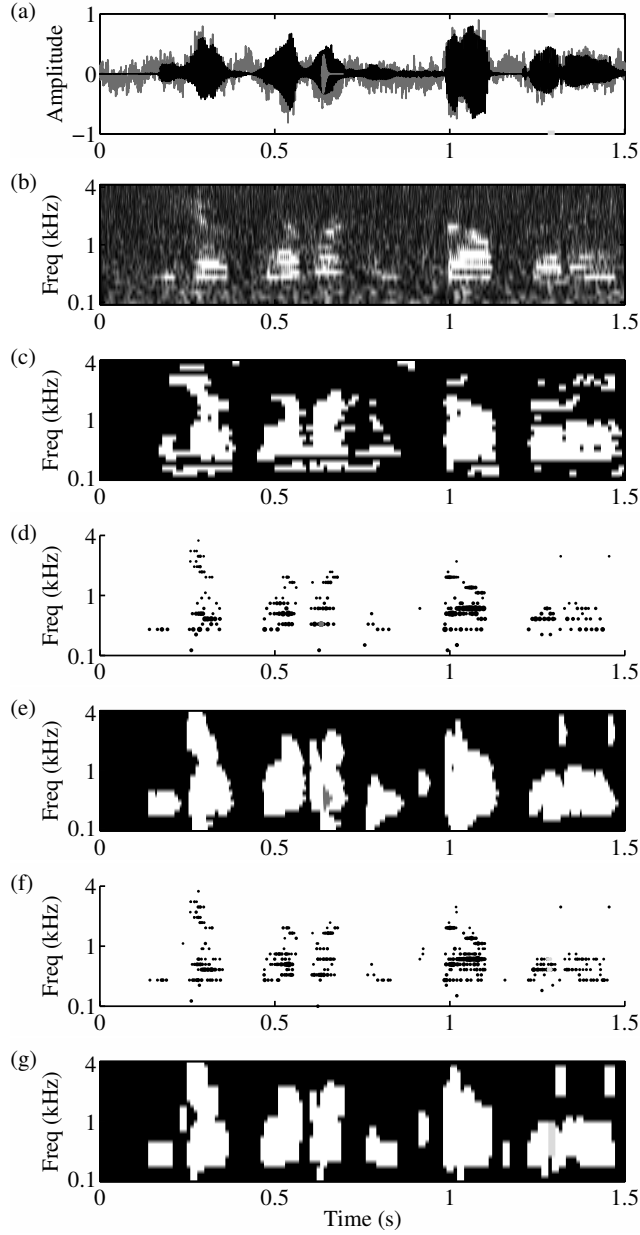


Figure 1: (a) Speech waveform (noisy in gray with the clean superimposed in black), (b) the gammatonegram of the noisy waveform, (c) the IBM, (d) the spikegram for MP with a single coefficient isolated in dark gray, (e) the MPBM with the corresponding T-F region for the dark gray coefficient in dark gray, (f) the spikegram for fMP with the coefficients for a single frame isolated in light gray, and (g) the fMPBM with the corresponding T-F regions in the isolated frame. (a) shows the gammatone that the dark gray MP coefficient corresponds to as well as the segment of the noisy waveform from which the light gray fMP coefficients are computed.

spikegrams contain coefficients in remarkably similar T-F regions. To illustrate the connection between fMP coefficients and the fMPBM, a single frame of coefficients is isolated in the spikegram in Fig. 1f with the color light gray. The corresponding segment of speech is designated on the waveforms in Fig. 1a with light gray lines along the horizontal axes as well. Finally, the T-F region that corresponds to the union of the contribution from each of the coefficients in the frame is highlighted in Fig. 1g with light gray. Since frames are overlapping, the mask is set to one if any of the associated frames designate it as target-dominated.

2.3.3 Filter-threshold binary mask

Although MP is theoretically amenable to real-time and low-power implementation, it is still highly nonlinear, and therefore, requires a nontrivial level of computation in order to choose the optimal set of coefficients. To gauge the necessity of this more optimal approximation, a third mask estimator is proposed that is based on FT (Smith and Lewicki, 2005). For filter and threshold binary mask (FTBM), filter response magnitudes for the mixture signal are computed, and the mask is given a value of one if the response magnitude is greater than or equal to the threshold.

2.4 Methods

Speech samples were created using the TIMIT database (1990) testing set re-sampled to 8 kHz. A male- and female-spoken sentence was chosen from each of the eight dialect regions to form a set of sixteen target sentences. To create the mixtures, pink noise, speech-shaped noise (SSN), and realistic noises from the AURORA-2 database (2000) were added as interferers at 0 dB and -5 dB SNR.

Noise reduction was performed on each of the mixtures using T-F gain manipulation with IBM and each of the binary mask estimation algorithms. For each strategy, the mixtures were decomposed with a gammatone filterbank containing 24 4th-order filters spaced one equivalent rectangular bandwidth (ERB) apart between 100 Hz and 4 kHz, each with a bandwidth of one-ERB (Hohmann, 2002). Mask estimation was performed in the T-F domain, and then the masks were applied point-wise to the filter response of the mixtures. For reconstruction, each frequency band of the modified mixture T-F representation was delayed and scaled so that the peaks of the impulse response of each band had a maximum at 4 ms. All of the frequency bands were then added together to obtain the processed waveform.

For the IBM method, the filter response magnitudes were computed for the target and interferer signals, and the energy was summed in each band across 20 ms time frames (Hamming window with 50% overlap). For each band, the mask was set to one at each of the 160 time samples within the frame if the target energy was greater than or equal to the interferer energy (i.e., 0 dB threshold).

For the MPBM and fMPBM methods, a dictionary of gammatones was created with impulse responses that matched those of the filterbank. Then for MPBM, the

MP coefficients are computed for the entire mixture at once, and for fMPBM, the MP coefficients are computed sequentially on 20 ms frames of the mixture with 50% overlap. For both, the approximation program is terminated when coefficients fell below a set of frequency-specific thresholds. To encourage coefficients in bands where speech energy is lower, the higher frequency bands had lower coefficient thresholds. Specifically, there was a linear scaling function (with a gain of four) between the lowest and highest frequency bands so that coefficients had an amplitude of at least 0.4 in the lowest band, at least 0.2 in the middle band, and at least 0.1 in the highest band.

To prevent a rapidly fluctuating mask in MPBM (particularly in the higher-frequency bands where the gammatone filter responses are very short in duration), each channel of the masks were post-processed with 10 ms frames (50% overlap) so that if the mask was initially set to one during any of the time samples in the frame, the final mask was to one at all time samples in the frame.

2.5 Results

2.5.1 Effect of frame-based processing

To explore the effect of frame-based processing on the MP-based algorithm, Fig. 2 shows the hit rates (true positive rate), false alarm rates (false positive rate), the hit-minus-false-alarm (H-FA) metric (the difference between the true positive rate and the false positive rate), and accuracy rates of MPBM and fMPBM using the threshold that yielded the maximum H-FA. For 0 dB SNR, thresholds of 0.67, 0.83, 1.00, and 1.00 yielded the maximum H-FA for pink noise, car noise, SSN, and babble, respectively. For -5 dB SNR, thresholds of 1.00, 0.83, 1.33, and 1.17 yielded the maximum H-FA rates for pink noise, car noise, SSN, and babble, respectively.

Based on H-FA, fMPBM is performing at similar levels as MPBM. In Kim et al. (2009), H-FA above 52% corresponded to significant increases in the percent of words normal-hearing listeners correctly identified in speech corrupted with babble, factory, and SSN at 0 dB and -5 dB SNR. With pink noise, both MPBM and fMPBM reach H-FA at or above 60% for both 0 dB and -5 dB input SNR levels. Performance decreases for the more challenging cases of speech corrupted with car noise, babble, and SSN. In these cases, the frame-based approach actually yielded higher H-FA averages than the non-causal approach because the false alarm rates decreased slightly.

2.5.2 Influence of sparse approximation algorithm

Fig. 3 shows predicted intelligibility outcomes for speech masked with the MPBM and FTBM approaches using the short-time objective intelligibility (STOI) measure, which was designed by Taal et al. (2011a) to maintain high correlation with subjective intelligibility of noisy and T-F-masked noisy speech. To compute the

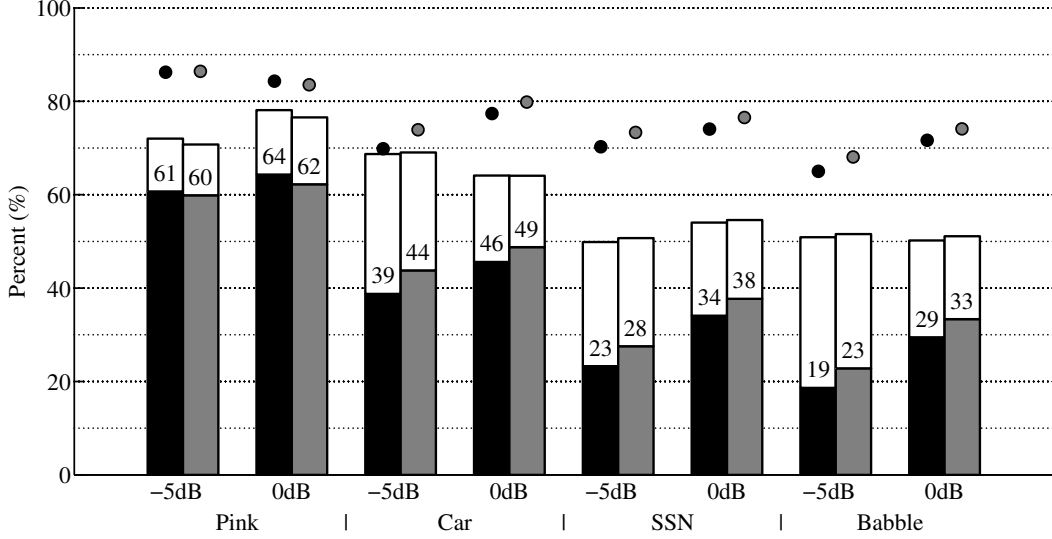


Figure 2: H-FA analysis of MPBM (in black) and fMPBM (in gray) for each of the noise types and input SNR levels. The filled sections of the bars indicate (with exact rates indicated above each bar); the non-filled sections of the bars indicate the false alarm rates; the tops of the bars indicate the hit rates; and the dots indicate overall accuracy.

STOI, the procedure from Taal et al. (2011a) was followed. Specifically, a degraded signal and its clean reference signal were decomposed into one-third octave bands and the envelopes were computed in each channel. The non-speech segments were removed. Then normalization was applied to compensate for global level differences, and clipping was applied to limit the influence of individual T-F units. The correlation coefficient was computed in each channel between the degraded and clean envelopes over short-time (384 ms) segments, and the average of the correlation coefficients was taken across all time and frequency for the signal. Finally, these values were converted to predicted percent correct intelligibility scores using the database-specific mappings as provided in Taal et al. (2011a).

STOI predictions for the IBM are consistent with results in the literature. However, STOI does not predict MPBM to increase intelligibility in the same way. For the case of pink noise, STOI predicts a small but insignificant increase in intelligibility when the MP threshold is around 0.4 for 0 dB SNR signals and when the MP threshold is around 0.6 for -5 dB SNR signals. With realistic noise, MPBM maintains intelligibility when the MP threshold is below about 0.25 for both -5 dB and 0 dB. In contrast, STOI predicts that intelligibility will degrade when the FTBM is applied in the presence of realistic noise. In the case of pink noise however, STOI predicts FTBM to maintain intelligibility at very low thresholds but never improve it.

Given that few practical algorithms exist for estimating the IBM in real-world applications, MPBM is a step in the right direction. However, STOI did not predict

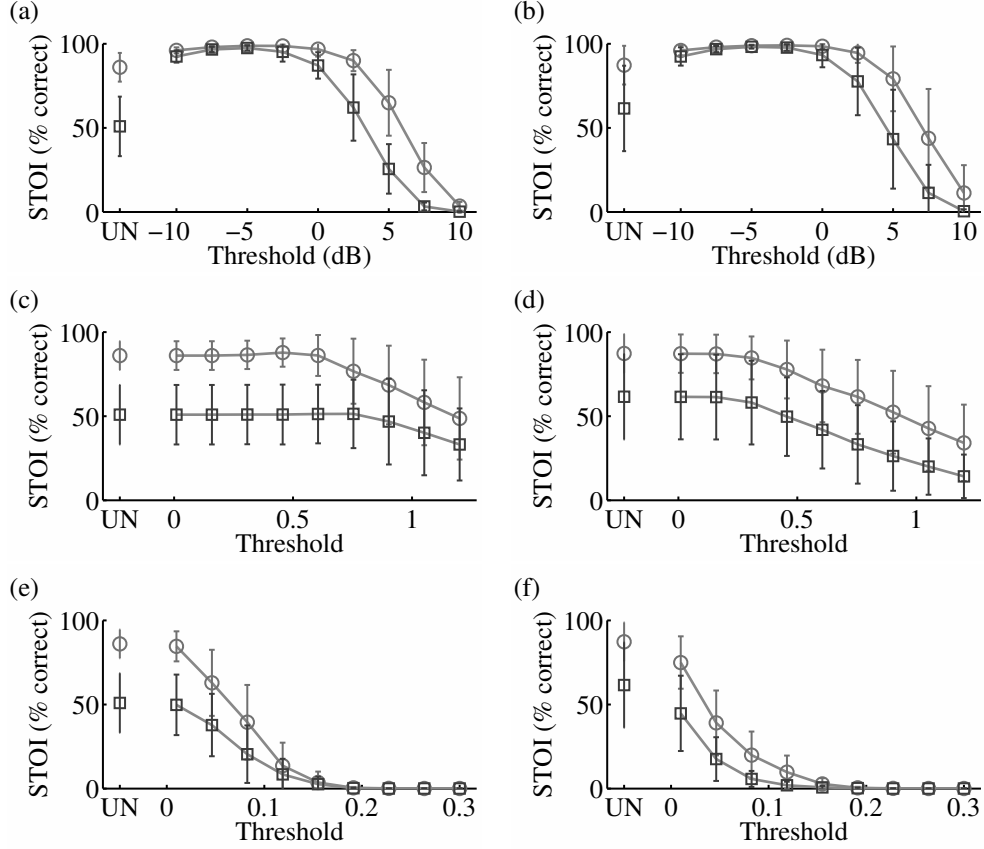


Figure 3: Average predicted intelligibility in percent words correct using STOI when the masks are applied to speech corrupted by pink noise [left column] and realistic noise (babble, car, street, and train) [right column] at 0 dB [circles] and -5 dB [squares] for a range of thresholds: (a-b) IBM, (c-d) MPBM, and (e-f) FTBM. indicates the unity mask or unprocessed condition. Error bars indicate plus and minus one standard deviation.

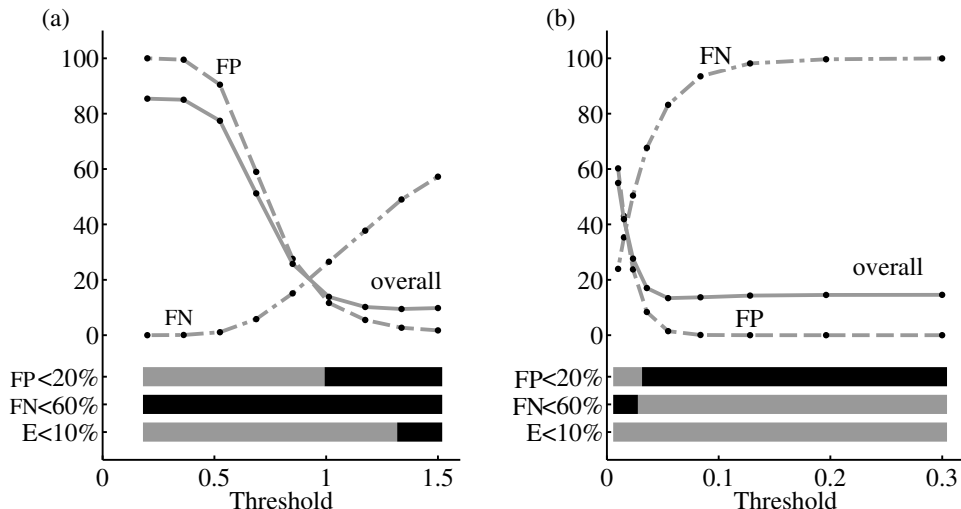


Figure 4: Average false positive (FP) error rates, false negative (FN) error rates, and overall error rates for (a) MPBM and (b) FTBM using a range of thresholds. The thick bars near the bottom indicate when each of the criteria are [black] satisfied and [grey] not satisfied.

that the current MPBM algorithm would increase intelligibility. To investigate MPBM further, it is necessary to look more specifically at the actual errors in the mask. Fig. 4 shows the average error rates for a range of MPBMs and FTBMs thresholds using sixteen sentences corrupted with pink noise at -5 dB. Taking into account the implied criteria by Li and Loizou (2008) and assuming that meeting all the criteria jointly leads to high intelligibility, a mask that satisfies all three criteria would be expected to yield high intelligibility. The FT algorithm fails to meet all three criteria jointly across the entire range of thresholds. Therefore, the poor intelligibility that STOI predicts is unsurprising. However, the MP algorithm with a threshold at or around 1.3 meets all three criteria jointly, and therefore, it is unexpected that STOI does not predict an intelligibility improvement.

2.6 Discussion

The preliminary results for fMPBM are encouraging, especially considering that this algorithm is amenable to an efficient implementation, it is causal, it does not require training, and it does not require oracle knowledge. Moreover, this study has demonstrated that MP can identify target-dominated T-F regions on a frame-by-frame basis just as well as it can on entire signals. However, performance for both

MPBM and fMPBM clearly suffers under certain noise conditions.

For comparison, conventional noise reduction algorithms such as the Wiener algorithm and the Ephraim and Malah suppression rule achieve H-FA on the order of -3% to 8% for speech corrupted by babble, factory, and SSN at -5 dB SNR when they are reformulated as T-F masking algorithms (Kim et al., 2009). Contrastingly, Han and Wang (2012) (classification using SVMs and auditory segmentation) report H-FA in the range of 36% to 59% for speech corrupted by babble, factory, and SSN at -5 dB and 0 dB SNR, and Wang and Wang (2012) (classification using CRFs and DNNs) report H-FA in the range of 75% to 82% for speech corrupted by noises at -10 dB, -5 dB, and 0 dB SNR. Although these reported numbers are higher than the numbers reported with the proposed MP-based approaches, it is difficult to be certain about how they compare to each other without a direct and more systematic comparison.

Broadly speaking, this study shows that sparsity-based approaches can do reasonably well, at least in terms of H-FA, when the model appropriately captures the statistical differences between the target and interferer, and since there is already progress towards solving sparse approximations with real-time, low-power architectures, sparsity-based methods are worthy of further consideration.

In an attempt to address the unsatisfying STOI result that the MP-based binary mask estimation algorithms are not expected to improve intelligibility, Fig. 5 looks more closely at the errors in the masks themselves for an example sentence rather than just the *rate* of errors. Three masks are compared against IBM with a 0 dB threshold: (a) IBM with a -10 dB threshold, (b) MPBM with a 1.3 threshold, and (c) a simulated mask with uniformly distributed random error similar to the artificially generated masks in Li and Loizou (2008). Visually, it is easy to see that the first mask contains many more target-dominated labels than the reference IBM, which leads to a very high “false positive” rate. With regard to the differences between the second and third masks, the errors in the second mask are grouped together in clusters, whereas the errors in the third mask are completely unstructured.

Again assuming that the IBM with a 0 dB threshold is ground truth, error rates were computed for each of the masks (Table 1). First, even with an overall “error rate” (relative to IBM with a 0 dB threshold) much higher than 10% , IBM with a -10 dB threshold yields very high intelligibility. Second, even though the error rates between the second and third masks are similar and both jointly satisfy the criteria suggested by Li and Loizou (2008), STOI predicts a reasonably high intelligibility for the case of randomly distributed errors, but very low intelligibility for the case of error that is grouped into clusters. Without listener studies, it is unclear how this clustering actually impacts the resulting intelligibility of the binary-masked speech.

To further explore the idea that structured errors influence intelligibility differently than randomly distributed errors, another example is given in the last row of Table 1. According to the study by Li and Loizou (2008) with randomly distributed

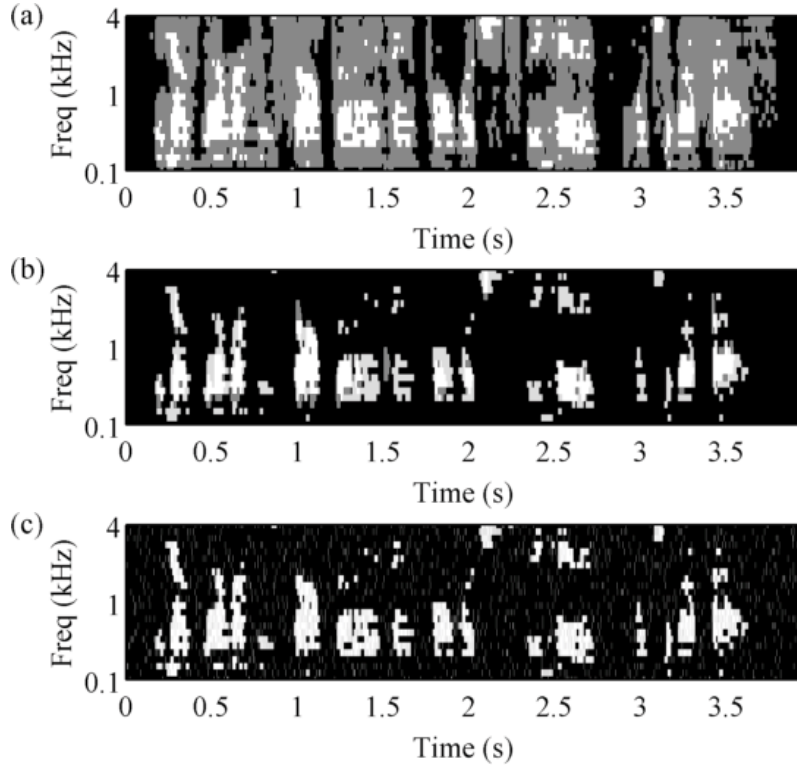


Figure 5: T-F representations of false positives (dark gray), false negatives (light gray), correct positives (white), and correct negatives (black) relative to the IBM with a 0 dB threshold for: (a) the IBM with a -10 dB threshold, (b) MPBM, and (c) a simulated mask with uniformly distributed random error.

Table 1: False positive (FP) error rates, false negative (FN) error rates, overall error rates, and STOI word recognition predictions for the masks in Fig. 5, as well as for IBM with a threshold of 2.5 dB.

Mask	Error rates			Intelligibility
	FP	FN	Overall	STOI
IBM with -10 dB threshold	55%	0%	47%	91%
MPBM	2%	59%	9%	8%
Mask with random errors	5%	26%	8%	74%
IBM with 2.5 dB threshold	0%	51%	7%	41%

error, it is expected that this mask would yield high intelligibility given that all three criterion on the error rates are met. Instead, STOI predicts low intelligibility. It seems that acceptable false positive and negative error levels are actually reversed for structured error that is grouped in T-F regions as compared to uniformly random error. To conclude, based on the analysis in this chapter, it is hypothesized that false negative errors are more tolerable when the error in question is randomly distributed, but false positive errors are more tolerable when the error in question is structured in T-F groups.

CHAPTER III

A SPEECH CODING STRATEGY BASED ON SPARSITY

Speech recognition is poor for cochlear implant recipients in listening environments with interfering noise. This study investigates the speech recognition in interfering noise that is provided by a coding strategy based on the sparse approximation algorithm matching pursuit. The speech recognition outcomes provided by the matching pursuit strategy is compared to that of traditional strategies as well as a strategy based on the ideal binary mask. The strategies are evaluated with an objective measure and with a listener study using vocoded speech and normal-hearing listeners. Despite the objective measure predicting a substantial improvement in speech recognition with the proposed strategy, word recognition scores in noise were actually lower with the new strategy compared to traditional strategies. Until more specific design criteria for channel selection strategies are constructed, it will be difficult to develop a realistic channel selection strategy that improves speech recognition in noise.

Parts of this chapter are based on the paper that was published as “Speech understanding in noise provided by a simulated cochlear implant processor based on matching pursuit” by Abigail A. Kressner and Christopher J. Rozell in the Proceedings of the *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, October 2013.

3.1 Introduction

For CI recipients, speech perception is poor in listening environments with interfering noise (Dorman et al., 1998). The poor outcomes are due in part to the fact that CIs provide limited frequency resolution, weak temporal pitch cues, a small dynamic range, and severely degraded temporal fine structure (Qazi et al., 2013). With these limitations in place, it is especially important to encode information in speech with a strategy that is both intelligent and efficient.

Two strategies that remain in widespread use in current CI systems are continuous interleaved sampling (CIS) and the Advanced Combination Encoder (ACETM) stimulation strategy (Wilson and Dorman, 2008). Studies have shown that CI recipients generally prefer ACE over CIS (Skinner et al., 2002). In noise however, the peak amplitude criterion that is at the root of ACE’s n -of- m channel selection algorithm can be problematic since it will prioritize encoding channels with high envelope amplitudes even if the channels contain only interfering noise (Hu and Loizou, 2008). To address this issue, Hu and Loizou (2008) proposed a selection criterion based on the IBM: channels with a local SNR greater than or equal to 0 dB are selected for stimulation. They demonstrated restoration of intelligibility for noisy speech when compared to CIS in quiet. However, as with IBM-based noise reduction, this selection criteria requires oracle knowledge of the target and interfering signal.

In line with the effort to employ the sparse coding model (see Section 2.2) to estimate the IBM, Taal et al. (2012) proposed a coding strategy based on MP. In their preliminary evaluation using tone-vocoded speech to simulate a CI processor, they demonstrated that their strategy improves intelligibility relative to the ACE strategy, as measured by objective measures, for speech in quiet. The main contribution of this chapter is to evaluate this MP-based channel selection strategy in quiet and in noise using both objective measures and listener studies with normal-hearing listeners.

3.2 Algorithms

CIS and ACE remain in widespread use within CI processors and are often the basis for comparison. In the CIS strategy, signals are decomposed into a small number of bands and the envelopes are extracted from each band. Variations in the envelope amplitudes are represented at corresponding electrodes in the CI through proportional modulation of trains of biphasic electrical pulses. The pulse trains for each of the channels and electrodes are interleaved in time so that the pulses across channels are non-simultaneous (Wilson and Dorman, 2008).

The ACE strategy belongs to the more general class of n -of- m algorithms. In n -of- m coding strategies, input signals are passed through m bandpass filters, and from these m bandpass filters, n channels are selected for stimulation. In ACE, the channels with the n largest envelope amplitudes are selected (Wilson and Dorman,

2008). Studies have shown that most users prefer the ACE over the CIS strategy (Skinner et al., 2002). However, while the peak amplitude criterion works well in quiet, it can introduce confusion in the presence of dominating noise because it prioritizes encoding channels with high envelope amplitudes even if the channels contain only interfering noise (Hu and Loizou, 2008).

To address the shortcomings of the ACE strategy, Hu and Loizou (2008) proposed using SNR for selecting which n channels to stimulate. This coding strategy is strongly based on the IBM; it selects channels with local SNRs that are larger than or equal to 0 dB and discards channels with local SNRs less than 0 dB. Unlike the n -of- m strategy which selects n channels to stimulate in every cycle, the IBM strategy chooses as few as zero channels and as many as m channels each cycle. When in quiet, the IBM strategy is equivalent to CIS. With six post-lingually deafened CI recipients, Hu and Loizou (2008) demonstrated nearly full restoration of speech intelligibility in noise when compared to CIS in quiet. However, as with all IBM-based processing this coding strategy requires prior knowledge of the target and interferer signals.

In contrast to IBM, the MP strategy (Taal et al., 2012) does not use prior knowledge. Instead, it uses a sparse coding model (see Section 2.2) to efficiently encode signals. Taal et al. (2012) reformulate CI channel selection as an inference problem in the T-F domain. They use a dictionary of complex exponentials characterized by frequencies that align with the CI electrodes. However, since inference is done on envelope amplitudes in the T-F domain rather than the time-domain waveforms as in Chapter 2, the dictionary is actually made up of the short-time filterbank magnitude responses to the set of windowed complex exponentials. Taal et al. (2012) also propose a weighting scheme that is derived from the STOI measure (Taal et al., 2011a). The weighting scheme is incorporated into MP to emphasize and de-emphasize channels in order to encourage or discourage selection of the corresponding dictionary elements. At each stimulation cycle, the MP strategy chooses the n channels that correspond to the first n dictionary elements that the weighted-MP program identifies as optimal.

3.2.1 Implementation

Let $x(t)$ denote the acoustic waveform, let $x_m(t)$ denote the m^{th} windowed segment of $x(t)$, and let $X_m(i)$ denote the short-time gammatone filterbank magnitude for the i^{th} band of $x_m(t)$ (van de Par et al., 2005). For CIS and IBM, the analysis rate and filterbank size match the stimulation rate and number of CI channels. Therefore, the stimulation pattern at the c^{th} cycle and the j^{th} channel $E_c(j)$ is given as a

function of $X_m(i)$ for the corresponding m^{th} frame and i^{th} frequency band as

$$E_c^{\text{CIS}}(j) = \sqrt{X_m(i)} \quad (2)$$

$$E_c^{\text{IBM}}(j) = \begin{cases} \sqrt{X_m(i)}, & \text{if } \text{SNR}_m(i) \geq 0 \text{ dB} \\ 0, & \text{if } \text{SNR}_m(i) < 0 \text{ dB}, \end{cases} \quad (3)$$

where $\text{SNR}_m(i)$ is the local SNR. For these algorithms, signals are divided into overlapping windows of 8 ms (Hann window with 50% overlap) and a stimulation pattern is recorded every 4 ms.

For MP, let B_m be a block of $X_m(i)$ over all I bands for the last M frames,

$$B_m = \text{vec} \left(\begin{bmatrix} X_{m-M+1}(1) & \cdots & X_m(1) \\ \vdots & \ddots & \vdots \\ X_{m-M+1}(I) & \cdots & X_m(I) \end{bmatrix} \right) \quad (4)$$

and let A_m be a set of channel-specific weights that are inversely proportional to the variation within each of the channels.

$$A_m = \text{diag} \left(\text{vec} \left(\begin{bmatrix} \frac{\sqrt{M}}{\sigma_m(1)} & \cdots & \frac{\sqrt{M}}{\sigma_m(1)} \\ \vdots & \ddots & \vdots \\ \frac{\sqrt{M}}{\sigma_m(I)} & \cdots & \frac{\sqrt{M}}{\sigma_m(I)} \end{bmatrix} \right) \right), \quad (5)$$

where $\sigma_m(i)$ is the standard deviation of the vector $[X_{m-M+1}(i) \cdots X_m(i)]$. For this strategy, the analysis windows are 16 ms in length (Hann window with 50% overlap), which is an analysis rate that is half as low as the stimulation rate (a stimulation pattern is generated every 4 ms) so two stimulation patterns are generated for each analysis frame. Thus, two different sets of dictionaries are utilized which are characterized by how the stimulation cycle aligns with the analysis window (see Taal et al., 2012, Fig. 1). The first channel to be selected in the stimulation pattern at the c^{th} cycle corresponds to the dictionary element that can represent the largest amount of energy in B_m . Put another way, MP chooses the dictionary element D_j with coefficient a_j that minimizes what the residual signal will become.

$$\hat{a}_j = \text{argmin} \|A_m R_x - a_j A_m D_j\|, \quad (6)$$

$$a_j = \frac{\langle A_m B_m, A_m D_j \rangle}{\|A_m D_j\|^2} \quad (7)$$

$$E_c^{\text{MP}}(j) = \sqrt{\hat{a}_j} \quad (8)$$

where R_x is the part of B_m not yet represented by stimulation in previous cycles (i.e., the residual). After selecting \hat{a}_j and recording it to $E_c^{\text{MP}}(j)$, the residual is

updated and the process is repeated until MP has identified the optimal set of n channels.

For all of the strategies, $J = 20$ channels are logarithmically spaced between 150 Hz and 5 kHz, and for MP and ACE, $n = 9$ channels are selected each cycle. The analysis filterbank is made up of $I = 20$ gammatone filters with center frequencies linearly spaced on an ERB scale between 150 Hz and 5 kHz for CIS, ACE, and IBM, and $I = 32$ gammatone filters linearly spaced between the same range for MP. Furthermore, $M = 48$ to form the T-F block B_m , which equates to approximately 400 ms.

3.2.2 MP strategy modifications

Fig. 6a-b shows the gammatonegrams for an example sentence in quiet and in SSN at 0 dB SNR. Fig. 6c-d show the corresponding electrodograms (a visualization of the stimulation pattern over time) using the MP strategy. In quiet, the electrodogram tracks the spectral peaks well, and there are clear boundaries between speech segments and gaps across all channels. In SSN however, MP channel selection tends to favor the highest channels.

The set of weights $\{\sqrt{M}/\sigma_m(i)\}$ as they evolve for the noisy sentence in Fig. 6b are depicted in Fig. 7a. At the onset of a large spectral peak, the standard deviations of associated channels increase significantly, and they remain high for the 400 ms that the peak is within B_m . Since the weights are inversely proportional to the standard deviations, these channels are de-emphasized or inhibited. Furthermore, since MP is consequently less likely to choose dictionary elements that encode the inhibited channels, the electrodograms tend to exhibit periods of low stimulation after large spectral peaks. Although Taal et al. (2012) designed this weighting scheme because of its correspondence with STOI and not because they specifically wanted to encode spectral peaks and gaps accurately, the weighting scheme tends to reproduce the clear boundaries that exist between strong speech segments and the subsequent gaps; even in noise, almost completely white regions follow the dark black regions in the electrodogram in Fig. 6d. Qazi et al. (2013) looked in detail at the effect of noise on stimulation strategies and intelligibility and concluded that preserving these clear boundaries is the most important factor for preserving intelligibility in noise.

Although the analysis thus far has identified some of MP's very promising characteristics, the strategy also has a couple of weaknesses. First, MP favors higher channels, and second, MP selects n channels even during speech gaps in quiet. Borrowing inspiration from IBM, MP would likely benefit from being able to recruit more or less channels when appropriate. To address these issues, two modifications are proposed to define the matching pursuit modified (MPm) strategy: (1) clip large weights so that the weighting scheme only produces inhibition (i.e., eliminate channel emphasis), and (2) let MP recruit as many channels as it wants as long as the

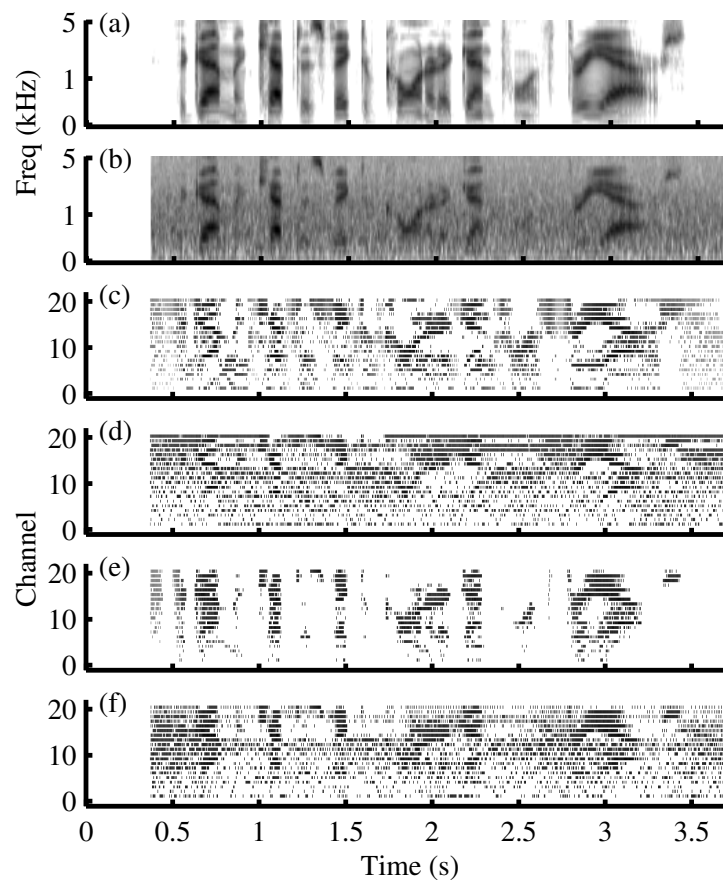


Figure 6: (a) Spectrogram of a sentence in quiet from the TIMIT database (1990): *the gunman kept his victim cornered at gunpoint for three hours*; (b) spectrogram in SSN (0 dB SNR); (c) electrodogram in quiet using the MP strategy; (d) electrodogram in SSN using the MP strategy; (e) electrodogram in quiet using the MPm strategy; and (f) electrodogram in SSN using the MPm strategy. The color map is on a logarithmic scale with black indicating large magnitudes, white indicating small magnitudes, and a range of 50 dB.

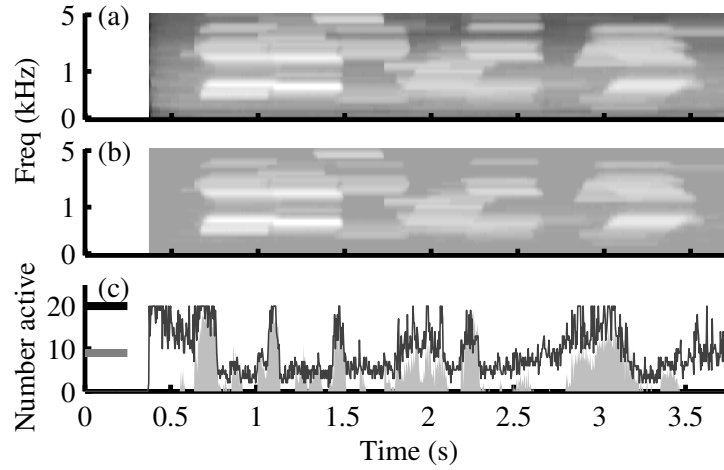


Figure 7: (a) MP weights for the noisy sentence shown in Fig. 6b (color map is on a logarithmic scale with black indicating large magnitudes, white indicating small magnitudes, and a range of about 80 dB); (b) MPm weights; and (c) the number of active channels during each stimulation cycle (CIS: black dash, MP: medium gray dash, IBM: light gray shaded area, MPm: dark gray thin line).

contribution to reducing the residual is large enough.

Fig. 7b shows the set of modified weights, which were heuristically chosen to be limited to 0.004. This modification substantially increased the robustness of the weights to noise, and as a result, reduced over-activity in the highest stimulation channels (Fig. 6f compared to 6d). Fig. 7c shows a comparison between of the number of active channels during each stimulation cycle of CIS, IBM, MP, and MPm. In CIS, the number of active channels is fixed to twenty, and in MP, the number of active channels is fixed to nine. In IBM, the number of active channels fluctuates with the number of channels that are target-dominated. When MPm is allowed to select all channels within a cycle that reduce the residual by at least 0.01% relative to the previous iteration, MPm tends to recruit additional channels during the same segments as IBM despite not having oracle knowledge.

3.3 Methods

3.3.1 Objective analysis

The CI coding strategies were evaluated first with an objective measure and then with listener studies using vocoded speech. For the objective analysis, 100 randomly selected sentences from the TIMIT database (1990) were mixtured at 0 dB with SSN and at 5 dB with babble noise from the AURORA-2 database (2000). The stimulation patterns were generated for the quiet and noisy versions using CIS, IBM, MP, and MPm, and then vocoded stimuli were synthesized. To synthesize

the vocoded stimuli, sinusoidal carriers were initialized with frequencies that match the center frequencies of the stimulation channels. The carriers were divided into 8 ms frames (Hann windowed with 50% overlap), modulated with the stimulation patterns, and then summed together. To predict intelligibility outcomes for MP and MPm in quiet and in noise as compared to CIS and IBM, an objective measure called normalized covariance metric (NCM) was implemented. NCM is an intelligibility measure based on the covariance between the envelopes of the test signal and its reference (the un-vocoded sentence in quiet) in each frequency band, and it has been shown to correlate highly with vocoded speech (Goldsworthy and Greenberg, 2004, Chen and Loizou, 2011).

3.3.2 Listener study

For the listener study, the speech material consisted of sentences from the IEEE database (1969) spoken by a female. Seventy-two lists of ten sentences make up the corpus. Each participant heard one randomly selected list for each mixture type and coding strategy, and none of the lists were repeated. The stimuli were noise-vocoded instead of tone-vocoded since noise-vocoders with channel overlap serve as a more effective tool for modeling the performance of CI listeners (Litvak et al., 2007).

Eighteen self-reported normal-hearing listeners participated in this study. All participants were native speakers of American English and students at the Georgia Institute of Technology. They were paid for their participation. At the start of the session, participants were instructed to listen through Sennheiser HD 202 headphones in a quiet room to a series of sentences. For each sentence, they were instructed to type the words that they heard into the response box using a graphical user interface (GUI). Participants were told that some of the sentences would be difficult to understand, but that a response must be made for each. Furthermore, they were told that typographical errors mattered and that guessing was okay. They could listen to each sentence only once and were allowed to take as many breaks as they wanted. To familiarize with the task, participants first listened to a set of ten sentences mixed with SSN at 10 dB using the ACE coding strategy. During this period, they were also asked to adjust the volume to a loud but comfortable level. After training, they were instructed not to change the volume for the duration of the session.

Speech reception thresholds (SRTs) were then measured for both SSN and babble noise using the QuickSIN test (Killion et al., 2004). Mixtures were subsequently created with SSN and babble noise at SNRs that were 1 dB lower than the SNRs of each participants estimated SRTs. Then the stimulation patterns were generated using ACE, IBM, and a parameter-optimized MPm. In the final phase of the session, fixed-level testing was conducted for each condition under comparison. In total, participants took between 1.5 and 2.5 hours to complete the test.

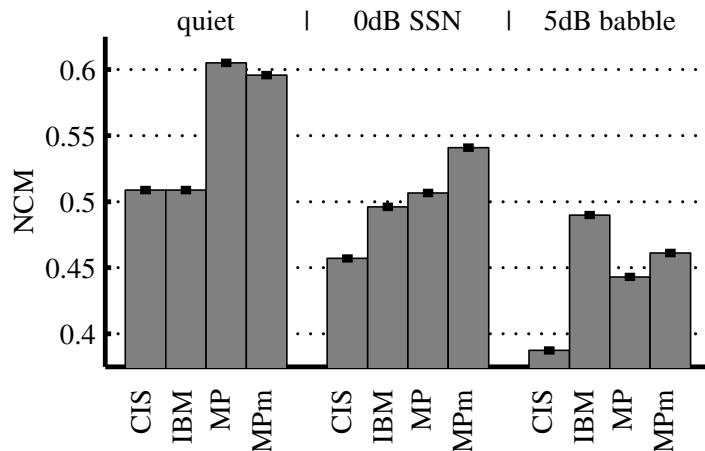


Figure 8: Mean NCM scores of vocoded signals as a function of the coding strategy for sentences in quiet, SSN (0 dB SNR), and babble (5 dB SNR). Standard error bars drawn in black.

Performance was evaluated by comparing every word transcribed by a participant against the set of words in the original sentence. The scoring was done automatically using a simple word recognition scheme. The participant’s transcription was first separated into words. If a word in the original sentence matched exactly one of the transcribed words, it was marked correct. Transcribed words were removed from the search once it was matched in order to avoid double-counting. Because of this automation, miss-spelled words, homophones, and compound words were not marked correct. This automatic scoring technique was verified for accuracy against a hand-scoring method and results were very highly correlated.

3.4 Results

3.4.1 Objective analysis

Fig. 8 shows the mean NCM scores for the tone-vocoded sentences. A three-way ANOVA found significant main effects of the coding strategy, interferer type, and sentence, as well as significant two-factor interactions ($p < 0.01$ for all). A multiple comparisons test with a Bonferroni adjustment (to control the family-wise error rate) over the interferer and coding strategy factors (i.e., population marginal means are computed for each combination of these two factors while removing the effects of the other factor) was also performed. All pairs were significantly different except CIS in quiet compared to IBM in quiet, both CIS and IBM in quiet compared to MP in SSN, CIS in SSN compared to MPm in babble, and IBM in SSN compared to IBM in babble.

NCM predicts that IBM almost fully restores intelligibility in both steady-state and fluctuating noise to the level predicted in quiet, which aligns well with the

results obtained in CI recipients (Hu and Loizou, 2008). In quiet, NCM predicts that MP will significantly improve intelligibility (which supports previous results in Taal et al. (2012)), and although the mean score for MPm is slightly lower, it also yields high scores. In SSN, NCM predicts that MP will fully restore intelligibility, and with MPm, it will actually increase intelligibility as compared to that of CIS in quiet. For both MP and MPm, the means are greater than the mean for IBM—a surprising result given that IBM utilizes oracle knowledge. In babble, NCM predicts that MP will significantly improve intelligibility as compared to CIS in babble, but that it only partially restores intelligibility as compared to CIS in quiet. Again, the modifications to the MP strategy yield significantly higher NCM scores. Unlike in the case of steady-state noise, MP and MPm do not outperform IBM in fluctuating noise.

3.4.2 Listener study

Fig. 9 shows the mean word recognition results for all participants. Two-way ANOVA with repeated measures indicated a significant effect of the mixture type ($F[2, 34] = 99.20, p < 0.001$) and a significant effect of the coding strategy ($F[2, 34] = 173.04, p < 0.001$). It also indicated a significant effect of the interaction between mixture type and coding strategy ($F[4, 68] = 26.55, p < 0.001$). A multiple comparisons test with Bonferroni adjustment over the estimated marginal means of the coding strategies for each mixture type indicated that the estimated marginal means for the three coding strategies significantly differed from each other ($p < 0.05$) for all except ACE and IBM in quiet and ACE and MPm in babble. In general terms, mean word recognition scores using the MPm strategy in its current form were lower than ACE in all noise types. These results differ drastically from the outcome that was predicted by the objective analysis with NCM.

3.5 Discussion

The listener study results confirm that an IBM-based channel selection strategy nearly restores speech recognition in noise to the level of speech recognition in quiet, which supports the results that were first reported in Hu and Loizou (2008). Unfortunately however, the stimulation patterns produced by the sparsity-based strategy do not appear to approximate the IBM stimulation patterns well enough to yield the same improvements in speech intelligibility. It is unexpected though that the NCM-based analysis predicted such high intelligibility outcomes for the MP strategy whereas the listener study yielded such low intelligibility scores for the MP strategy. Further investigation is needed in order to fully understand the poor outcome of this listener study with the proposed sparsity-based channel selection strategy. Such an investigation would likely consist of an analysis of, for example, when the channel selection strategy incorrectly picks a channel that was not supposed to be stimulated or fails to pick a channel that was supposed to be stimulated. This type

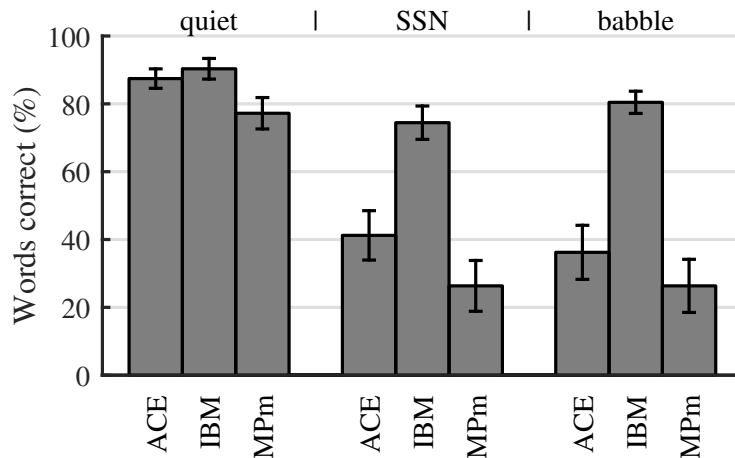


Figure 9: Mean word recognition scores of vocoded speech as a function of the coding strategy for sentences in quiet, SSN, and babble. Confidence intervals drawn in black.

of analysis is similar to the type of analysis used in the machine learning community to evaluate binary mask estimation algorithms. Unfortunately, in the same way as binary masks, very little information is available about how errors impact intelligibility when the errors are made in a structured way. As a result, it is challenging to design new strategies because there is a lack of guidance about which information is important to encode correctly and when.

Although vocoded speech is an imperfect model of CI recipient outcomes, vocoded speech intelligibility at least demonstrates how much of the spectral information is represented in a coding strategy’s output. Given the poor word recognition scores of the vocoded speech encoded with the proposed sparsity-based strategy, further development is warranted before a sparsity-based channel selection strategy can be recommended for testing in a CI processor. However, taking into consideration the crudeness of the n -of- m strategies and the low levels of speech recognition for CI recipients in noise, there is much room for improvement in CI channel selection. Based on the fact that the channel selection problem is well-described by the sparse coding model, it remains a promising direction.

CHAPTER IV

STRUCTURE IN TIME-FREQUENCY BINARY MASKING ERRORS AND ITS IMPACT ON SPEECH INTELLIGIBILITY

This chapter describes a novel investigation framework for binary masks and presents listener studies with normal-hearing listeners that use this framework to illustrate how interactions between error types and structure affect intelligibility. First, this study demonstrates that clustering of mask errors reduces intelligibility. Furthermore, while previous research has suggested that false positives are more detrimental to intelligibility than false negatives, this study indicates that false negatives can be equally detrimental to intelligibility when they contain structure or when both error types are present. Finally, this study shows that listeners tolerate fewer mask errors when both types of errors are present, especially when the errors contain structure.

This chapter is based on the paper to be published as “Structure in time-frequency binary masking errors and its impact on speech intelligibility” by Abigail A. Kressner and Christopher J. Rozell in the *Journal of the Acoustical Society of America*, accepted March 2015.

4.1 Introduction

Although studies of independent and uniformly random mask errors provide some useful information about error tolerance, they provide little information about how interactions between error types and structure in mask errors affects intelligibility performance. This uncertainty provides motivation for a systematic listener study in this direction. The main contributions of this chapter are to present an investigation framework for binary masks and to present listener studies that use this framework to illustrate the effect of more realistic mask errors. Generating masks with specific amounts of false positives and false negatives while containing specific levels of T-F structure is nontrivial, and the framework presented includes the development of a statistical model for mask errors that can generate examples with error structure similar to that observed in mask estimation algorithms. Using this model to generate masks with specific amounts of error and structure, three listener studies were conducted. Generally, the first two studies mimic that of Li and Loizou (2008) by presenting varying amounts of false positives alone or false negatives alone, respectively, but they also explore how structure in the errors influences intelligibility outcomes. The third listener study subsequently investigates the intelligibility outcomes of the more realistic scenario in which masks contain interacting errors (i.e., simultaneous false positives and false negatives) with and without structure.

4.2 Description of the model

A graphical model is a probabilistic model that captures the conditional dependence between random variables (Bishop, 2006). Fig. 10 illustrates the form of a model for binary masks that captures the error rates of a non-ideal binary mask relative to the IBM as well as the dependencies between adjacent T-F units in the non-ideal mask. The form of this model is similar to the well-studied *Ising model* for image denoising (Bishop, 2006, pg. 389). Each node in the bottom layer of the model corresponds to the T-F units in a non-ideal mask (i.e., the latent random variables x_i), and the edges between those nodes represents the conditional dependence between neighboring T-F units. Each node in the bottom layer is also connected via an edge to a node in the top layer. These nodes in the top layer correspond to the T-F units in the IBM (i.e., the observed random variables y_i), with the edges representing the probability that the non-ideal T-F unit takes the same state as the IBM unit (i.e., the false positive and negative error rates). While richer statistical models could capture more expansive definitions of structure in binary mask estimates, the simplest model that was sufficient to capture essential characteristics in estimated masks was chosen to avoid excess complexity.

This graphical model can perform two functions. First, it can randomly generate non-ideal binary masks (i.e., *model-generated masks*) from an IBM using a technique called *sampling*. Second, it can “learn” the conditional dependencies within a set of estimated binary masks using a technique called *parameter estimation* (also known

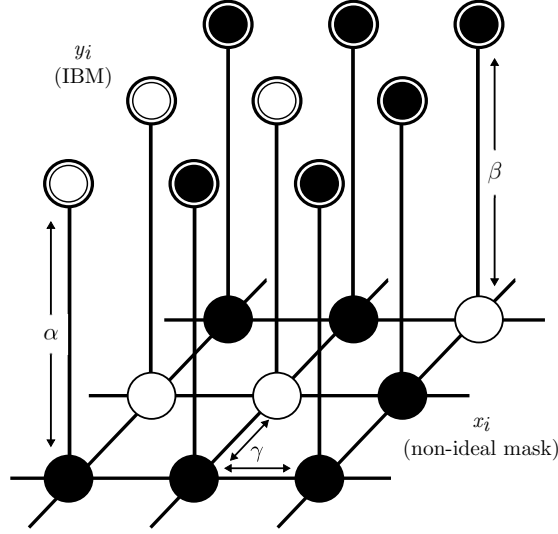


Figure 10: Illustration of a graphical model for binary masks that captures the false positive error rate (α) and the false negative error rate (β) in a non-ideal mask relative to the IBM, as well as the statistical dependence between adjacent T-F units in the non-ideal mask (γ). In the sampling framework, the non-ideal mask is a model-generated mask, whereas in the parameter estimation framework, the non-ideal mask can also be an estimated binary mask.

as *training*). In order to describe the techniques of sampling and parameter estimation in more detail, it is necessary to define the probability distribution $p(\mathbf{x}, \mathbf{y})$ for a pair of masks first.

Mathematically, $p(\mathbf{x}, \mathbf{y})$ describes the joint distribution over the entire set of x_i and y_i . The probability of a particular assignment of all the variables can be represented as a normalized product of a set of non-negative potential functions

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \prod_i \phi_i(x_i) \prod_i \phi_i(y_i) \prod_{\langle i, j \rangle} \phi_{\langle i, j \rangle}(x_i, x_j) \cdots \prod_{\{i, j\}_0} \phi_{\{i, j\}_0}(x_i, y_j) \prod_{\{i, j\}_1} \phi_{\{i, j\}_1}(x_i, y_j), \quad (9)$$

where Z is the normalizing constant that forces the distribution to sum to one and ensures that the model defines a valid probability distribution. The potential functions $\phi_i(x_i)$ and $\phi_i(y_i)$ give a non-negative weight to each possible value of the random variable. The notation $\langle i, j \rangle$ indicates all the non-ideal node pairs i and j that are nearest neighbors, $\{i, j\}_0$ indicates all the non-ideal and ideal node pairs i and j where the ideal node is interferer-dominated, and $\{i, j\}_1$ indicates all the non-ideal and ideal node pairs i and j where the ideal node is target-dominated. Finally, the potential functions $\phi_{\langle i, j \rangle}(x_i, x_j)$, $\phi_{\{i, j\}_0}(x_i, y_j)$, and $\phi_{\{i, j\}_1}(x_i, y_j)$ give a non-negative weight to each possible combination of values of the pair of associated random variables.

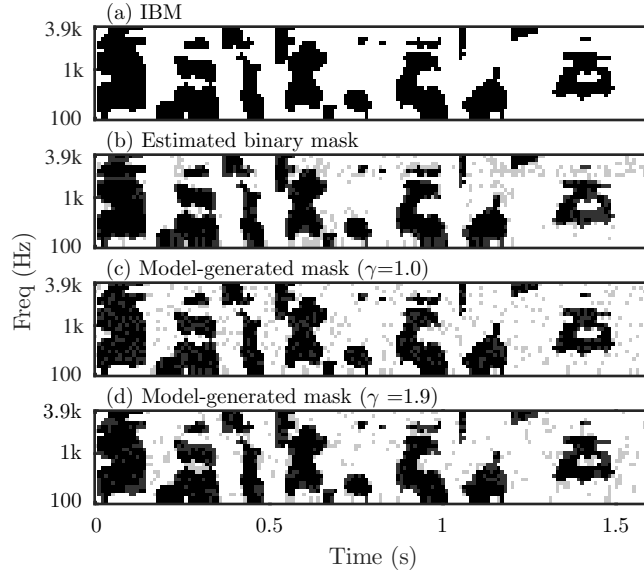


Figure 11: (a) An IBM for the example sentence, “Bad nerves are jangled by a door slam,” mixed with babble noise at 5 dB SNR, (b) an example estimated mask using a GMM-based classifier, (c) a mask with uniformly random, uncorrelated errors, and (d) a model-generated mask with correlated errors ($\gamma = 1.9$). The false positive and false negative errors in the three non-ideal masks are marked with light gray and dark gray, respectively. Each of these masks contain the same error rates ($\alpha = 11\%$ and $\beta = 27\%$) but the errors differ in structure. Note that parameter estimation with the model yields $\gamma = 1.9 \pm 0.1$ for the GMM-based estimated masks.

4.2.1 Parameterization

The potential functions in a graphical model are parameterized by a set of model parameters θ . Depending on the formulation of a particular graphical model and the complexity of the parameterization, θ can vary in size. To motivate the design of the parameterization for the graphical model in this work, Fig. 11 presents a few example binary masks. Fig. 11a shows the IBM for a sentence mixed at 5 dB SNR with babble noise from the AURORA-2 database (2000), and Fig. 11b shows an example GMM estimated mask for the same sentence. Throughout this chapter, the GMM-based approach is used to exemplify authentic masks as this type of processing has been applied repeatedly in the literature (e.g., Kim et al., 2009, Kim and Loizou, 2010a,b, Hu and Loizou, 2010, Han and Wang, 2012, May and Dau, 2013). In the figure, the false positive and false negative errors are marked with distinct gray tones to differentiate between the errors. For visual comparison, Fig. 11c shows a mask that contains uniformly random errors (i.e., completely unstructured) with the same amount of false positive and false negative errors as the GMM mask. Note that these unstructured errors are the same kind of errors that Li and Loizou (2008) introduced to the masks in their experiments.

Qualitatively, the errors in the GMM binary mask are significantly more clustered than the errors in the mask with unstructured errors, showing the presence of substantial T-F structure in the errors of the GMM algorithm. To quantitatively study the effects of this clustering on how much mask error listeners can tolerate in binary-masked speech, masks need to be generated in a way that allows the clustering and the amount of false positives and false negatives to be varied systematically. Carefully tuning actual estimation algorithms to do this would be tedious and arguably unfeasible because algorithm parameters do not usually map directly and independently to false positive and negative rates nor to the amount of clustering. However, a strategic parameterization of the model will enable this kind of tuning in model-generated masks. To accomplish this, θ consisted of the parameters A , B , γ , and λ . Specifically, $\phi_{\{i,j\}_0}(x_i, y_j) = \begin{bmatrix} 1-A & A \end{bmatrix}$ so that A effectively controlled the bias towards mislabeling interferer-dominated units; $\phi_{\{i,j\}_1}(x_i, y_j) = \begin{bmatrix} B & 1-B \end{bmatrix}$ so that B effectively controlled the bias towards mislabeling target-dominated units; $\phi_{\langle i,j \rangle}(x_i, x_j) = \begin{bmatrix} \gamma & 1 \\ 1 & \gamma \end{bmatrix}$ so that γ effectively controlled the bias towards neighboring units in both the temporal and spectral directions having the same label rather than having different labels; and $\phi_i(x_i) = \begin{bmatrix} \lambda & 1-\lambda \end{bmatrix}$ and $\phi_i(y_i) = \begin{bmatrix} \lambda & 1-\lambda \end{bmatrix}$ so that λ effectively controlled the bias towards choosing interferer-dominated labels.

Since the values of these potentials cannot be interpreted in exact probabilistic terms, a set of more intuitive probabilistic error characteristics were also defined: the false positive rate α (related monotonically to A) and the false negative rate β (related monotonically to B). Consequently, mask error configurations can effectively be characterized by the set of α , β , and γ . To give more intuition about the clustering parameter, changing the parameter to $\gamma = 2.0$, for example, indicates that neighboring T-F units are twice as likely to be in the same state as they are to be in different states, and $\gamma = 1.0$ indicates that the errors are unstructured with a uniformly random distribution. Therefore, a mask with $\gamma = 2.0$ will contain more clustering than a mask with $\gamma = 1.0$.

4.3 Methods

The model presented in the previous section is employed to generate stimuli for three experiments conducted in three separate sessions. The first experiment investigated the influence of structure in false positive binary mask errors; the second experiment investigated the influence of structure in false negative binary mask errors; and the third experiment investigated the influence of interacting unstructured and structured false positives and negatives. To determine which levels of clustering were most relevant to evaluate in these studies, parameter estimation was first performed on estimated masks to assess the level of clustering in prevalent authentic masks.

4.3.1 Assessment of realistic clustering levels

The representative estimation algorithm was a reduced version of the GMM-based estimation algorithm presented in May and Dau (2013). With this algorithm, a set of sample estimated and ideal mask pairs were generated and then parameter estimation was performed on the set of mask pairs. The process consisted of the following steps: (1) initializing the mixtures that would make up the classifier-training and parameter-estimation data sets, (2) computing the amplitude modulation spectrogram (AMS) features for both data sets, (3) computing the IBMs for both data sets, (4) training the classifier with the classifier-training data set (i.e., learning a “map” between AMS features and the IBMs), (5) generating estimated masks for the mixtures in the parameter-estimation data set (i.e., using the “map” and AMS features to predict masks), and (6) estimating the parameters of the graphical model using the parameter-estimation data set of mask pairs (i.e., learning the resulting relationship between the estimated masks and the IBMs).

For the classifier-training data set, 180 sentences spoken by females were randomly selected from the IEEE database (1969), down-sampled to 16 kHz, and mixed at -5 , 0 , and 5 dB SNR with randomly selected segments of resampled babble from the NOISEX database (1993). For the parameter-estimation data set, 60 new randomly selected sentences were resampled and mixed at -5 dB SNR with randomly selected segments of resampled babble noise from the same respective databases. Computation of the AMS features and the IBMs, training of the classifier, and generation of the estimated masks were all completed using similar methods as the full version of the estimation algorithm by May and Dau (2013). However, this reduced version of the algorithm used only the AMS features and the two-class Bayesian classification stage and excluded the background noise classification stage, the pitch-based and speech presence probability features, and the T-F integration neighborhood function. To form the final set of mask pairs for the parameter estimation stage, the estimated and ideal masks were truncated in time so that all 120 of them were of equal dimensions.

For the purposes of parameter estimation, the task is to compute the potential functions that maximize the likelihood of a set of data. Given a conditional graph structure, parameterization θ of the model, and a set of data, the optimal values of θ are the set that minimize the negative log-likelihood function,

$$\arg \min_{\theta} (-\log p_{\theta}(\mathbf{x}|\mathbf{y})). \quad (10)$$

For binary masks, this is equivalent to finding the parameters that best describe the error characteristics produced by the set of example estimated binary masks from the representative algorithm. Generally, the minimization can be solved using gradient descent or quasi-Newton methods, and to compute the negative log-likelihood function at each iteration during minimization, inference (i.e., estimation of Z and the marginal probabilities) is performed as a sub-routine.

To conduct the estimation, a graphical model was constructed for all 60 instances of the estimated masks and each T-F unit was assigned two binary indicator features designating which of the two labels in the associated IBM unit was the correct label. Then, all of the weights in the potential functions were tied to the value of one or to one of a set of free weights depending on the type of potential function and its position within the mask. The potential functions were then computed by exponentiating the assigned weights. The minimization was performed using the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm, which approximates Newton’s method to find a stationary point of the objective function (Liu and Nocedal, 1989). The inference was executed as a sub-routine using the mean field method, which approximates the full graphical model using a simpler graphical model that has a probability distribution with minimal Kullback-Leibler divergence from the original model but that makes the computation of the marginals easier. After conducting the optimization program, the amount of clustering in the estimated masks was, by definition, the value of the edge potential between x_i nodes that are in the same state relative to the value of the edge potential between x_i nodes that are in different states.

Quantification of the clustering in the sample set of GMM masks was approximately $\gamma = 1.9$. This value was relatively robust to changes in the estimation algorithm parameters (e.g., the sampling rate, the number of filters, window length, the interferer type, the input SNR, etc.) since the value of γ varied between only 1.8 and 2.0 in all of the exploratory trials conducted with this estimation algorithm. Although a good way to quantitatively measure how accurately this model captures all of the structure in estimated mask errors does not exist, Fig. 11 provides an anecdotal comparison of the example GMM estimated mask and a model-generated one with the learned parameter $\gamma = 1.9$ (Figs. 11b and 11d). While it is evident from visual inspection that this model does not capture the full extent of the structure in estimated mask errors, the masks generated with $\gamma = 1.9$ contain errors that clearly capture more of the structure than uniform random errors (i.e., $\gamma = 1.0$).

Conceptually, this result means that neighboring T-F units in these realistic masks are approximately twice as likely to be in the same state as they are to be in different states. A γ value of 1.0 would have meant that this kind of structure does not exist in these sample masks. However, since γ was reliably greater than 1.0, it is confirmed that realistic masks do make errors that contain structure. Furthermore, this exercise provides an approximate range for which to investigate how clustering in mask errors influences intelligibility outcomes.

4.3.2 Stimuli

The first experiment evaluated word recognition at four levels of false positive rates ($\alpha = 10\%$, 20% , 40% , and 60%) with four levels of clustering ($\gamma = 1.0$, 1.5 , 2.0 , and 2.5). A control condition called the “unity mask” (un; i.e., a mask of all ones) condition was also included to make a total of 17 testing conditions. The

second experiment evaluated word recognition at four levels of false negative rates ($\beta = 10\%$, 20% , 40% , and 60%) with four levels of clustering ($\gamma = 1.0$, 1.5 , 2.0 , and 2.5). The un condition was also included to make a total of 17 testing conditions. The third experiment evaluated word recognition at three levels of false positive rates ($\alpha = 10\%$, 20% , and 40%) simultaneously with three levels of false negative rates ($\beta = 10\%$, 20% , and 40%) and with two levels of clustering ($\gamma = 1.0$ and $\gamma = 2.0$). The un condition was included again to make a total of 19 testing conditions.

The speech material consisted of sentences from the IEEE database (1969) spoken by a female and down-sampled to 8 kHz. The sentences were corrupted with randomly selected segments of babble noise from the AURORA-2 database (2000) (sampling rate of 8 kHz) at -5 dB SNR. The babble masker started 100 ms before the beginning of each sentence recording and continued for 100 ms after the recording. The first list in the IEEE database (1969) was isolated for task familiarization for all subjects, and the remaining 71 lists were used in testing. Subjects heard two randomly selected lists per condition (i.e., 20 corrupted sentences per condition), and none of the lists were repeated within a session. The lists were presented in separate blocks, and the first block of all of the conditions was presented before any of the second blocks. The order of the conditions within the first and second halves of the sessions was randomized across subjects.

For each mixture signal within a condition, the IBM was computed with a local target-to-interferer threshold of -5 dB, 32 gammatone filters with center frequencies spaced between 80 Hz and 3.7 kHz, and 20 ms segments (50% overlap) using the implementation by Wang et al. (2008). For all conditions except the un conditions, a graphical model was then configured for the IBM to produce the characteristics of the condition. Specifically, A was set according to the desired level of α ; B was set according to the desired level of β ; γ was set to the desired level of the clustering; and λ was set so that it was approximately 0.5 (the actual value varied ± 0.05 according to the content of each mixture but it was subsequently determined that a constant value would have been adequate). The values of A and B were tuned heuristically in advance for each combination of α , β , and γ to achieve averages of α and β across the collection of mixtures presented to all subjects at the desired rates. Note that the actual value of λ was of little importance because A and B were tuned heuristically to achieve the desired α and β relative to the IBMs.

After configuring the model, the values of y_i were fixed to the values of the IBM. By fixing the elements of \mathbf{y} to the observed binary values of the T-F units in the IBM, the model implicitly defines a conditional distribution $p(\mathbf{x}|\mathbf{y})$ over the non-ideal binary mask, and the model becomes a CRF – a type of discriminative undirected graphical model. For the purposes of sampling, the task is to randomly draw a set of values x_i from the posterior distribution $p(\mathbf{x}|\mathbf{y})$. In other words, sample configurations of the set of x_i are drawn according to the probability of them occurring given the configuration of the set of y_i and the parameterization θ . One sample configuration of the set of x_i forms a model-generated mask.

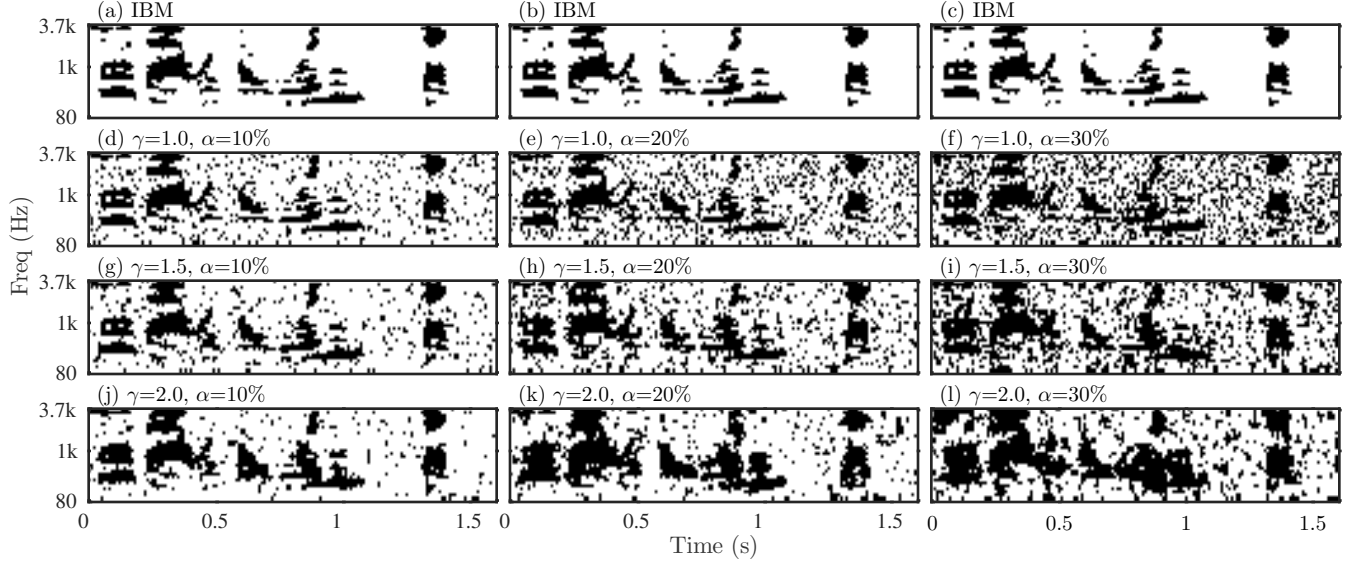


Figure 12: The top row contains three copies of the IBM for the example sentence, “Do that with a wooden stick,” mixed with babble noise at -5 dB SNR. The remaining rows contain model-generated masks with false positives and varying amounts of clustering. Specifically, the second row contains masks with $\gamma = 1.0$ (uniformly random, uncorrelated), the middle row with $\gamma = 1.5$, and the bottom row with $\gamma = 2.0$. The left column contains masks with $\alpha = 10\%$, the middle column with $\alpha = 20\%$, and the right column with $\alpha = 30\%$.

Model-generated masks were drawn from the conditional distribution using the Gibbs sampling algorithm, which is a Markov chain Monte Carlo algorithm for obtaining a sequence of observations which are approximated by the joint probability distribution (Bishop, 2006). Because samples from the beginning of the chain may not accurately represent the desired distribution, the first 100 steps (i.e., the *burn-in period*) are discarded, and the state of the chain on the 101th step is used as a sample of the joint distribution.

Finally, the mask was applied to the mixture, and the mixture was resynthesized. Fig. 12 shows nine model-generated masks containing only false positives with a range of clustering ($\alpha = 10\%$, $\alpha = 20\%$, and $\alpha = 30\%$ with $\gamma = 1.0$, $\gamma = 1.5$, and $\gamma = 2.0$). This figure illustrates how the conditions vary in the first experiment. Fig. 13 shows nine model-generated masks containing only false negatives with a range of clustering ($\beta = 10\%$, $\beta = 20\%$, and $\beta = 30\%$ with $\gamma = 1.0$, $\gamma = 1.5$, and $\gamma = 2.0$). This figure illustrates how the conditions vary in the second experiment. Finally, Fig. 14 shows six model-generated masks containing both false positives and negatives ($\alpha = 10\%, \beta = 10\%$; $\alpha = 10\%, \beta = 20\%$; and $\alpha = 20\%, \beta = 10\%$) and just two levels of structure ($\gamma = 1.0$ and $\gamma = 2.0$). This figure illustrates how the conditions vary in the third experiment.

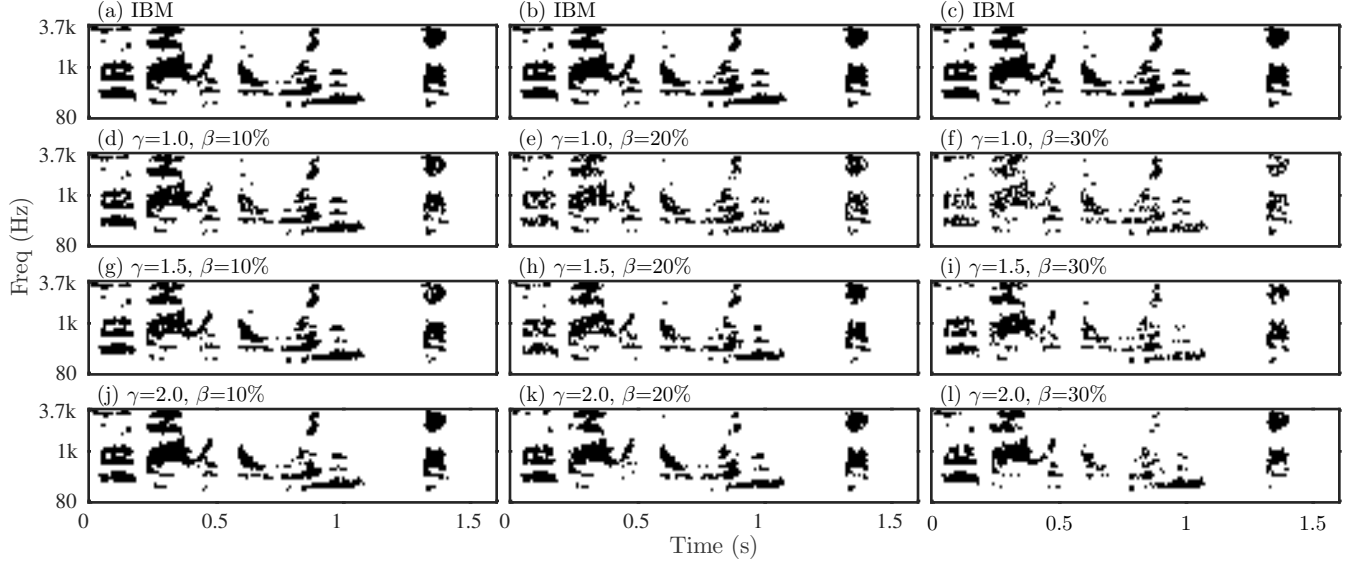


Figure 13: The top row contains three copies of the IBM for the example sentence, “Do that with a wooden stick,” mixed with babble noise at -5 dB SNR. The remaining rows contain model-generated masks with false negatives and varying amounts of clustering. Specifically, the second row contains masks with $\gamma = 1.0$ (uniformly random, uncorrelated), the middle row with $\gamma = 1.5$, and the bottom row with $\gamma = 2.0$. The left column contains masks with $\beta = 10\%$, the middle column with $\beta = 20\%$, and the right column with $\beta = 30\%$.

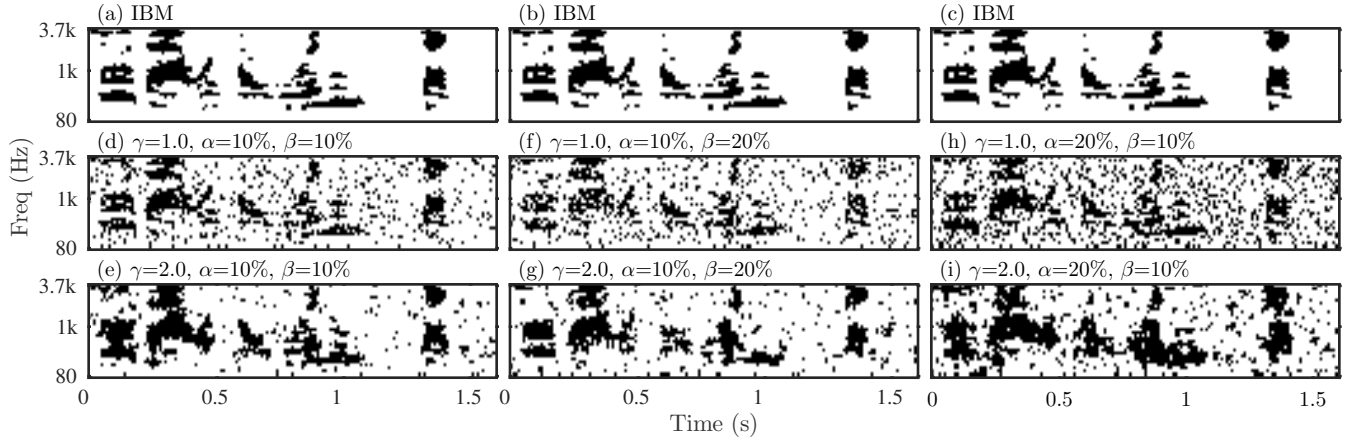


Figure 14: The top row contains three copies of the IBM for the example sentence, “Do that with a wooden stick,” mixed with babble noise at -5 dB SNR. The remaining rows contain model-generated masks with false positives, false negatives, and varying amounts of clustering. Specifically, the second row contains masks with $\gamma = 1.0$ (uniformly random, uncorrelated) and the bottom row with $\gamma = 2.0$. The left column contains masks with $\alpha = 10\%$ and $\beta = 10\%$, the middle column with $\alpha = 10\%$ and $\beta = 20\%$, and the right column with $\alpha = 20\%$ and $\beta = 10\%$.

4.3.3 Subjects

Ten self-reported normal-hearing listeners (five female and five male) participated in all three experiments. All subjects were native speakers of American English and were paid for their participation. The subjects' ages ranged from 22 to 29 years, and all were undergraduate or graduate students at the Georgia Institute of Technology.

4.3.4 Procedure

During each session, a GUI was presented to subjects using a laptop, and the subjects were instructed to listen through Sennheiser HD 202 headphones in a quiet room to a series of sentences. For each sentence, they were instructed to type the words that they heard into the response box. Subjects were told that some of the sentences would be difficult to understand, but that a response must be made for each. Furthermore, they were told that typographical errors mattered and that guessing was okay. They could listen to each sentence only once and were allowed to take as many breaks as they wanted. To familiarize with the task, subjects first listened to a set of ten mixture signals processed with the un control condition. During this period, they were also asked to adjust the volume to a loud but comfortable level. After training, they were instructed not to change the volume for the duration of the session. In total, subjects took between 1.5 and 2.5 hours to complete each of the experiments.

Each subject's performance was evaluated by comparing every word transcribed against the set of words in the original sentence. The scoring was done automatically using a simple word recognition scheme. The subject's transcription was first separated into words. If a word in the original sentence matched exactly one of the transcribed words, it was marked correct. Transcribed words were removed from the search once it was matched in order to avoid double-counting. Because of this automation, miss-spelled words, homophones, and compound words were not marked correct. This automatic scoring technique was verified for accuracy against a hand-scoring method and results were very highly correlated.

4.4 Results

4.4.1 Effect of structure in false positive errors

This listener experiment investigated the effect of structure on the tolerance of false positives. Fig. 15 shows the mean word recognition scores for all conditions as well as 95% confidence intervals. Two-way ANOVA with repeated measures indicated a significant effect of γ ($F[3, 27] = 54.35$, $p < 0.001$), a significant effect of α ($F[3, 27] = 177.75$, $p < 0.001$), and a non-significant interaction of γ and α ($F[4.55, 40.96] = 1.77$, $p = 0.147$). Because the result of the Mauchly's test of sphericity was significant for the interaction, a Greenhouse-Geisser adjustment was made to the respective degrees of freedom. Generally, increased clustering

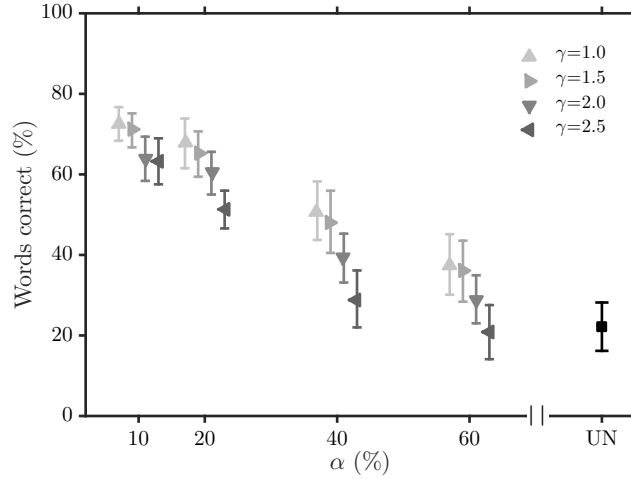


Figure 15: Mean word recognition scores and 95% confidence intervals for a range of error configurations with varying levels of false positive rates (α) and clustering structure (γ). un is the unity mask control condition.

resulted in increased sensitivity to false positive errors, especially when the error rates approached higher levels.

For the unstructured conditions ($\gamma = 1.0$), the overall trend between false positive error rate and word recognition in the study by Li and Loizou (2008) was replicated. However, group average word recognition was generally lower. The higher scores in the Li and Loizou study were likely due to the fact they used the exact same segment of noise for every mixture, they presented all 20 of their mixtures for each condition in the same block so subjects had more time to adjust to the processing, and they hand-scored their results so they counted homophones and compound words as correct.

4.4.2 Effect of structure in false negative errors

This listener experiment investigated the effect of structure on the tolerance of false negatives. Fig. 16 shows the mean word recognition scores for all conditions as well as 95% confidence intervals. Two-way ANOVA with repeated measures indicated a significant effect of γ ($F[3, 27] = 97.74$, $p < 0.001$), a significant effect of β ($F[3, 27] = 176.38$, $p < 0.001$), and a significant interaction ($F[4.01, 36.09] = 12.54$, $p < 0.001$). Because the result of the Mauchly's test of sphericity was significant for the interaction, a Greenhouse-Geisser adjustment was made to the respective degrees of freedom. Similar to the previous experiment, increased clustering resulted in increased sensitivity to false negative errors, especially when the error rates approached higher levels. The significance of the interaction provides evidence that the effect of false negatives depends on the level of γ .

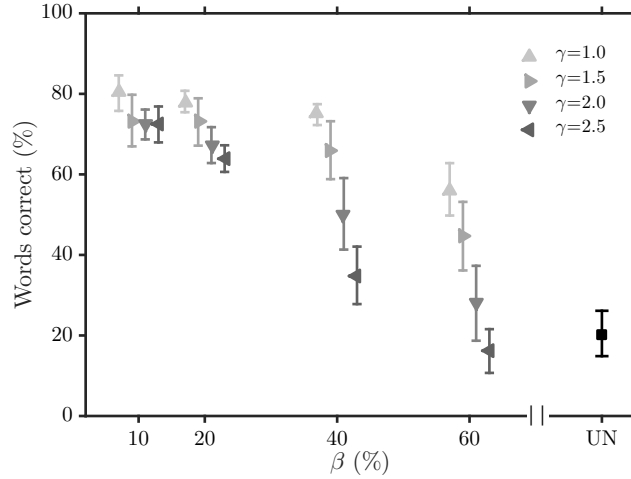


Figure 16: Mean word recognition scores and 95% confidence intervals for a range of error configurations with varying levels of false negative rates (β) and clustering structure (γ). un is the unity mask control condition.

For the unstructured conditions ($\gamma = 1.0$), the general plateau relationship between false negative error rate and word recognition in the study by Li and Loizou (2008) was replicated. Specifically, intelligibility was relatively resilient to increasing false negative errors when $\gamma = 1.0$. However, when the errors contained increasing amounts of clustering, the plateau behavior diminished and false negative errors became increasingly detrimental. Based on the observation that the plateau behavior diminished when the errors contained clustering and that higher levels of clustering in the errors yielded substantially lower intelligibility, it is clear that false negative errors can be just as detrimental to intelligibility as false positive errors when they contain structure. This result differs from that of the study by Li and Loizou (2008) and highlights the importance of accounting for structure in the errors.

4.4.3 Effect of interacting errors and structure

This listener experiment investigated how unstructured and structured false positives and negatives interact. Fig. 17 shows the mean word recognition scores for all conditions. Three-way ANOVA with repeated measures indicated a significant effect of γ ($F[1, 9] = 241.00, p < 0.001$), α ($F[1.30, 11.67] = 129.45, p < 0.001$), and β ($F[2, 18] = 155.88, p < 0.001$). Furthermore, it indicated a non-significant effect of the interaction between γ and α ($F[2, 18] = 0.11, p = 0.894$), a significant effect of the interaction between γ and β ($F[2, 18] = 5.23, p = 0.016$), and a non-significant effect of the interaction between α and β ($F[4, 36] = 2.14, p = 0.096$). Finally, the analysis also indicated a significant effect of the interaction between all three factors ($F[4, 36] = 3.59, p = 0.015$). Because the result of the Mauchly's test of sphericity was significant for the effect of α , a Greenhouse-Geisser adjustment was made to

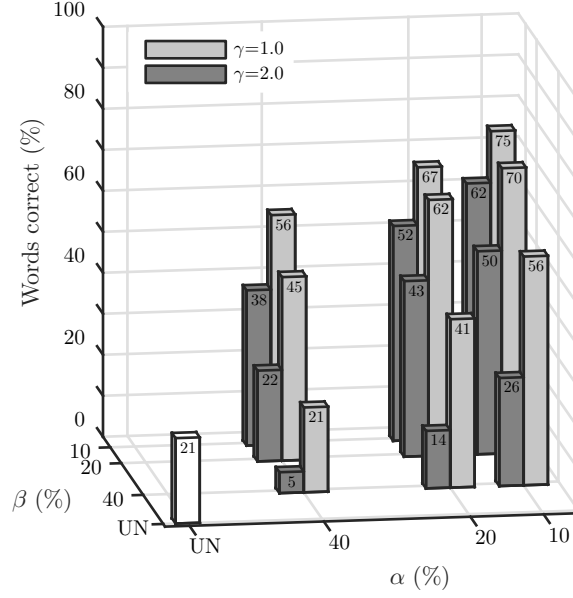


Figure 17: Mean word recognition scores for a range of error configurations with varying levels of false positive rates (α), false negative rates (β), and clustering structure (γ). un is the unity mask control condition.

the respective degrees of freedom.

Similar to the first and second experiments, increased clustering in the error configurations resulted in increased sensitivity to the errors, and all factors had a significant main effect. Additionally, the significance of the γ and β interaction in this experiment provides further evidence that the effect of the false negative error rate changes depending on the level of clustering. Contrastingly, given the lack of significance of the interaction between γ and α in this experiment and in the first experiment, the same is not true of the false positive error rate. Furthermore, there is no definitive evidence showing that the effect of the false positive error rate changes if the level of the false negative error rate changes. In spite of that, the significance of the three-way interaction provides evidence that the effect of false positive and negative error rates collectively depend on the structure.

Focusing first on the effect of structure in the errors, for almost all of the error configurations, the intelligibility outcome for the masks with clustered errors was significantly worse than the intelligibility outcome for the masks with the unstructured errors. By increasing the amount of the clustering of the errors, larger T-F regions of the interferer and target were added and removed at one time. As expected, adding and removing larger T-F regions of the signal negatively impacted intelligibility.

Turning next to the effect of interacting errors, the non-symmetric paired conditions were designed to explore the importance of each type of error when both were

present. Comparing $\alpha = 10\%, \beta = 20\%$ with $\alpha = 20\%, \beta = 10\%$; $\alpha = 10\%, \beta = 40\%$ with $\alpha = 40\%, \beta = 10\%$; and $\alpha = 20\%, \beta = 40\%$ with $\alpha = 40\%, \beta = 20\%$, intelligibility outcomes within each structure level were not significantly different between pairs with flipped α and β rates ($p < 0.05$ for all paired comparisons). Therefore, when both error types were present, false negatives were equally harmful to intelligibility as false positives. This result supports the conclusion in Section 4.4.2 that false negatives can be just as detrimental as false positives, but extends it to say that false negatives with *and without* structure can be as detrimental as false positives when both error types are present.

In addition to exploring the relative importance of each error type, this study was designed to explore whether or not the tolerances of false positives and false negatives change when both error types are present (in comparison to the tolerances when only one type is present). Specifically, the goal was to test whether or not jointly satisfying the implied criteria in the Li and Loizou (2008) study (i.e., α should be less than 20% and β less than or equal to 60%) would yield maximally high intelligibility. To focus on specific examples for clarity, average word recognition in the first experiment was greater than 70% when $\alpha = 10\%$, $\beta = 0\%$, and $\gamma = 1.0$ (Section 4.4.1) and average word recognition in the second experiment was greater than 70% when $\alpha = 0\%$, $\beta = 40\%$, and $\gamma = 1.0$ (Section 4.4.2). However, intelligibility was significantly lower at only 56% in this third experiment when $\alpha = 10\%$, $\beta = 40\%$, and $\gamma = 1.0$. Thus, jointly satisfying the criteria by the independent experiments did not necessarily yield maximally intelligible speech. This destructive behavior of interacting errors was present when the errors were unstructured, and it was even more substantial when the errors were clustered. Consequently, the effects of false positives and false negatives should be investigated simultaneously rather than, or at least in addition to, being investigated independently.

Fig. 18 shows contour plots of the mean word recognition scores for combinations of α and β relative to the mean word recognition score in the un condition. According to the plot, if the errors are unstructured ($\gamma = 1.0$), all combinations of α and β that fall on or below the solid line marked +20% would lead to intelligibility improvements of at least 20%. Contrastingly, if the errors are clustered ($\gamma = 2.0$), only combinations of α and β that fall on or below the dashed line marked +20% would lead to intelligibility improvements of that magnitude. Comparing the two, the data suggests that the restrictions on the combinations of α and β are much more rigorous when the errors are clustered. Note also that masks with clustered errors containing rates of errors in combinations of α and β that fall above the dashed line marked +0% would actually be detrimental to intelligibility. However, these contour plots are based on limited data, and further investigation is needed in order to make more conclusive guidelines about the tolerance of mask errors in realistic estimated masks.

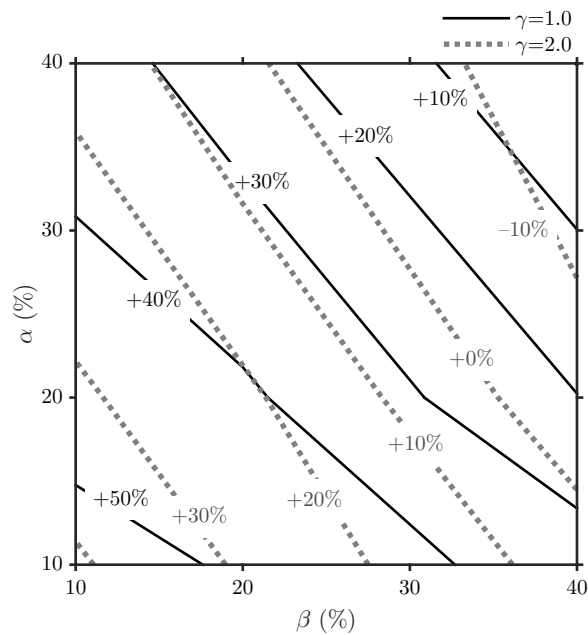


Figure 18: Data re-plotted from Fig. 17. Contours of the mean change in word recognition scores relative to the unity mask (UN) condition. When mask errors are unstructured ($\gamma = 1.0$), combinations of α and β on or below the solid contour lines indicate mean word recognition changes equal to or better than the indicated amount. In contrast, when the mask errors are clustered with $\gamma = 2.0$, combinations of α and β on or below the dashed contour lines indicate mean word recognition changes equal to or better than the indicated amount.

4.5 Discussion

Despite the encouraging capabilities of the simple method presented here, there are likely even more accurate ways to model the underlying structure in estimated binary masks. For example, adding an additional γ parameter to decouple the statistical dependence across frequency from the statistical dependence across time would likely improve the model's ability to represent the structure in speech onsets and transients more accurately. Alternatively, adding an additional variable to the model that *explicitly* captures the proximity to the edges between target- and interferer-dominated regions would likely improve the model's accuracy as well. All of these extensions are implementable in an altered form or parameterization of the graphical model and provide interesting future directions.

In addition to using the proposed model to generate synthetic masks, the model could also be used to analyze more carefully the errors that specific estimation algorithms tend to make or what underlying structure in the signal the algorithm fails to capture. Subsequent to this analysis or alongside it, the model could be used to help tune algorithm parameters by, for example, minimizing certain types of structure in the errors or maximizing the representation of certain types of structure in the target signal. To highlight the flexibility of graphical models in general, a graphical model could also be used to implement a particular stage within an estimation algorithm or even used as the primary classifier of an estimation algorithm (note that Wang and Wang (2012), for example, have proposed a method in this direction).

The GMM-based estimation algorithm that was presented as a representative algorithm in Section 4.3.1 did not include the T-F integration neighborhood function that May and Dau (2013) proposed to improve the classification performance of the original GMM algorithm. This neighborhood function was designed specifically to incorporate the information of adjacent T-F units into the classification decision so that, for example, the probability of a single T-F unit being target-dominated is dependent on the probability of the neighboring units also being target-dominated. The integration of information across neighboring T-F units drastically increased the performance of their estimation algorithm; however, this integration function is also likely to increase the amount of clustering in the resulting masks, and in turn, increase the requirements on the accuracy of the algorithm. This cyclical phenomena is of particular interest for developers of new algorithms and calls for more investigation into the extent to which different algorithms (or parts of algorithms) alter the amount of clustering. Along this same line, it will be important to investigate the implications of clustering on estimation algorithm evaluation metrics in the literature, particularly because the most commonly used metrics do not take underlying structure into account. Taken together, these investigations could considerably impact the direction of research for new estimation algorithms as well as impact the way in which new algorithms are evaluated.

These experiments only investigated and compared the effect of error rates and

structure in IBMs on speech intelligibility. So while a specific subset of processing parameters might yield similar intelligibility scores, the resulting speech quality could be vastly different. Therefore, it would be interesting to expand the current study to investigate the effect of different types of errors and structure on the quality of binary-masked speech. In general, sound quality and listener preference of IBM processing has been the topic of multiple studies already (e.g., Brons et al., 2012, Anzalone et al., 2006, Mauger et al., 2012). It is clear from these studies that because of the inherent processing artifacts, the IBM will likely never be used directly in hearing aids or cochlear implants without modification (although it is noteworthy to mention that the required modifications may be fairly moderate in cochlear implants). Nevertheless, IBM processing improves intelligibility. Therefore, understanding which specific aspects of the algorithm lead to improved intelligibility, and which are detrimental for quality, will facilitate development of noise reduction algorithms in the future.

To summarize, this chapter presented an investigation framework as well as listener studies that use the framework to characterize the influence of mask errors in more realistic configurations. The results from the three listener experiments collectively demonstrated the following:

1. **Clustering of binary mask estimation errors reduces intelligibility.**

The general trend between speech intelligibility outcomes and false positive and false negative error rates from the Li and Loizou (2008) study were replicated in the first two studies in the conditions with $\gamma = 1.0$. As the masks contained increasing levels of clustering in time and frequency ($\gamma > 1.0$), listeners were able to tolerate consistently fewer mask errors. This was especially true when both error types were present. Conceptually, this result means that adding in and removing larger T-F regions of the interferer and target signals (as opposed to adding in and removing randomly distributed individual T-F units) negatively influences word recognition of the target signal. Although the definition of structure in this study is simple, it is clear from the results that the underlying structure in mask errors can dramatically influence intelligibility outcomes.

2. **False negatives can be just as detrimental as false positives when the errors contain structure or when both error types are present.** This conclusion conflicts with the conclusion in the Li and Loizou (2008) study that false positives are more detrimental to intelligibility. Although it is true that false negatives have little impact when the errors are unstructured and there are no false positives, realistic algorithms will estimate masks that contain both types of error simultaneously with nontrivial amounts of structure. Thus, false negatives can be as influential on intelligibility as false positives in realistic scenarios.

3. **The rates of errors listeners can tolerate decreases when both types**

of errors are present. Jointly satisfying the implied criteria from the individual false positive and false negative studies does not necessarily yield high intelligibility. In other words, the listeners tolerated smaller rates of mask errors when both types of errors were present than when each was present separately. This is especially true when these errors contain structure. As a result, the effects of false positives and false negatives should be investigated simultaneously rather than or in addition to being investigated independently.

The proposed model is a useful tool for generating algorithm-independent masks with specific amounts of random error and structure. Even though the formulation of the model in this study captures only basic structure, it demonstrates that even this simple structure can substantially influence intelligibility outcomes. By exploring richer models of structure, this tool should be able to generate synthetic masks with even more realistic error types and, as a result, would facilitate more comprehensive investigations of how structure influences intelligibility. Furthermore, this tool could be used to characterize the errors that estimation algorithms tend to make and therefore may prove useful for tuning the parameters in these algorithms. All things considered, this study and the proposed investigation framework may help in the future to identify more specifically how and why the IBM successfully improves intelligibility, and ultimately, aid the development of algorithms that improve speech intelligibility in noisy environments without prior knowledge of the target and interferer signals.

CHAPTER V

COCHLEAR IMPLANT SPEECH INTELLIGIBILITY OUTCOMES WITH STRUCTURED AND UNSTRUCTURED BINARY MASK ERRORS

Although the effect of both unstructured and structured binary mask errors was investigated with normal-hearing listeners in Chapter 4, it has not been investigated with cochlear implant recipients. This chapter assesses speech recognition of binary-masked speech with cochlear implant recipients using binary masks with model-generated errors. The results demonstrate the following: (1) clustering of mask errors substantially decreases the tolerance of errors, (2) incorrectly removing target-dominated regions can be more detrimental to intelligibility than incorrectly adding interferer-dominated regions if the errors are clustered, and (3) the individual tolerances of the different types of errors decreases when both are present. These trends follow those of normal-hearing listeners. Although the basic trends are similar though, the data provides some indications that CI recipients may actually be more susceptible to mask errors.

5.1 *Introduction*

With the IBM-based channel selection strategy (Hu and Loizou, 2008) introduced in Chapter 3, CI recipients achieved speech recognition scores in noise that were nearly equivalent to their speech recognition scores in quiet with the traditional channel selection method, which is an increase in word recognition on the order of as much as 65%. Following this study, Mauger et al. (2012) showed that an IBM-based noise reduction (in combination with ACE) also substantially improves CI speech intelligibility outcomes.

Chapter 4 investigated the impact of structure, as well as interacting error types, on speech intelligibility with normal-hearing listeners. This chapter demonstrated firstly that the tolerance of each individual error type decreases when both types of errors are present at the same time. Secondly, it showed that clustering of binary mask estimation errors reduces the tolerance of the errors as well. Finally, it showed that false negatives can be just as detrimental to intelligibility as false positives when the errors contain structure or when both error types are present, thereby contradicting the conclusions of Li and Loizou (2008) that false positives are more detrimental.

However, the outcomes of CI recipients with T-F gain manipulation do not always mimic those of normal-hearing listeners (Mauger et al., 2012). In fact, Qazi et al. (2013) concluded that CI recipients tolerate less noise in the gaps between speech segments than their normal-hearing counterparts. Since false positive errors can introduce noise in speech gaps, it is worthwhile to determine whether this translates to a higher sensitivity to false positive errors in CI recipients as compared to normal-hearing listeners. Furthermore, Qazi et al. (2013) also concluded that modulations in the stimulation current level during speech segments have little to no effect on the intelligibility of speech as long as the channel selection remains ideal. In this way, the binary decision about when and where to stimulate is particularly important in the framework of CIs. Thus, a systematic study in this area is warranted in order to better understand how mistakes in this decision process impact intelligibility specifically for CIs. The main contribution of this chapter is to present a listener study that assesses speech intelligibility in CI recipients with binary-masked speech using artificially generated masks that contain specific error configurations and specific amounts of structure.

5.2 *Methods*

The model presented in Chapter 4 was employed to generate masks with specific error configurations, and with these masks, mixtures of target speech and interfering signals were preprocessed acoustically using a common binary-masking technique. The processed mixture signals were presented to CI recipients via a real-time computer system that mimics their own sound processors. Participants were scored on their recognition of words morphemically. Details about the model, the generation

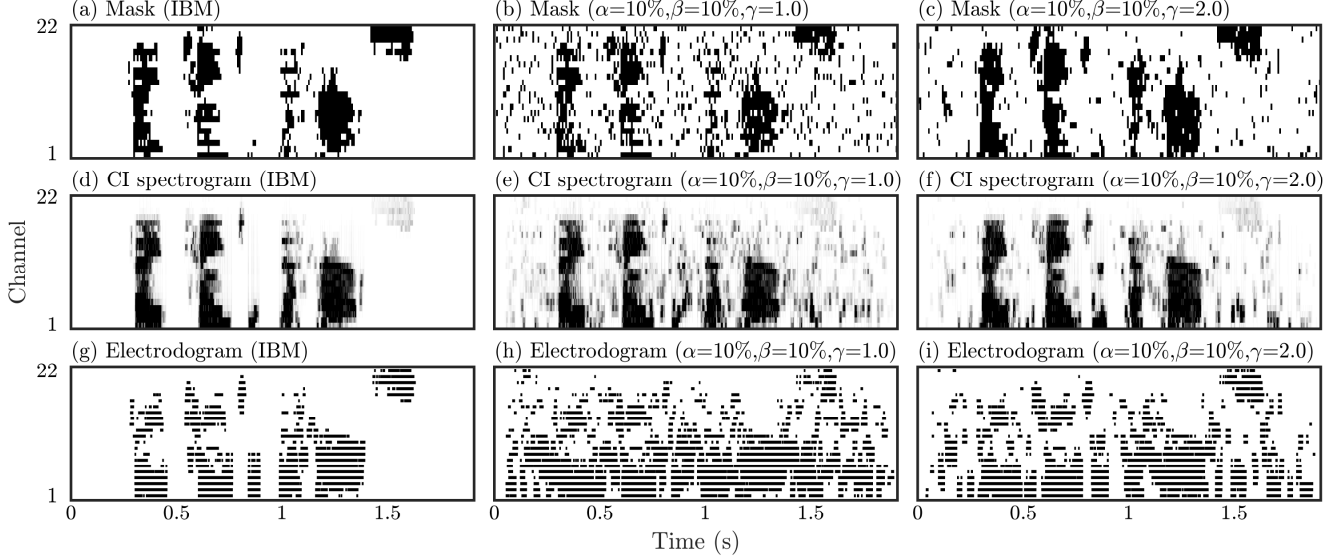


Figure 19: (a) Example ideal mask, (b) example model-generated mask with unstructured error, (c) example model-generated mask with structured error, (d-f) spectrograms from inside the cochlear implant model (using P1’s sound processor settings) of the sentence that was mixed at 2 dB with babble and then binary-masked with the masks in (a) through (c), respectively, and (g-i) the resulting stimulation sequences (i.e., electrodograms) for those sentences, respectively.

of stimuli, the CI participants, the setup of the system, and the testing procedure are given below.

5.2.1 Stimuli

The target speech material was taken from an Australian corpus of sentences that were designed similarly to the original Bamford-Kowal-Bench (Bench et al., 1979) sentences (i.e., the “BKB-like” corpus). The sentences were developed by the Cooperative Research Centre for Cochlear Implant and Hearing Aid Innovation (CRC HEAR). Eighty lists of sixteen sentences (1280 sentences in total) make up the corpus. Each sentence is comprised of four to six words or six to eight syllables using vocabulary that is familiar to a five-year-old, that is not specific to any particular region of Australia, and that was recorded by a female Australian English speaker at 44.1 kHz. The interfering noise material consisted of approximately 30 seconds of sixteen-talker babble that was constructed by superimposing two different segments of a single recording of eight-talker babble. The original eight-talker babble was a recording at 44.1 kHz of four Australian English females and four Australian English males speaking original Bamford-Kowal-Bench (BKB) sentences. The root-mean-square (RMS) levels of all individual sentences and the babble were equalized.

Each participant heard two randomly selected lists per condition, and none of

the lists were repeated. Each condition was presented once in the first half of the testing session and once in the second half of the experiment. The order of the conditions within the first and second halves was randomized across participants.

For each sentence, the mixture signal was generated by corrupting the target signal with a randomly selected segment of the interferer at every SNR from -5 to 10 in 1 dB increments. In each, the interferer started 200 ms before the beginning of the target sentence and continued for 200 ms after the end of the sentence, and the overall amplitude was gradually increased and then decreased during these segments using a cosine ramp. The signals were bandpass-filtered with a 1024 -point finite impulse response (FIR) filter designed with cutoff frequencies at 187.5 and 7937.5 Hz, and then re-equalized to have the same RMS level as the individual signals. Next, the mixture signal was segmented into overlapping frames of 720 samples (approximately 16 ms duration) with a shift of 360 samples. Each frame was windowed with a modified cosine window (i.e., a full cosine cycle multiplied by 0.5 and raised by 0.5), and 720 -point fast Fourier transforms (FFTs) were computed. A mask consisting of the binary weightings for each T-F unit was applied to the spectral-temporal representation of the mixture to realize the specific experimental condition. After the mask was applied, the waveform was resynthesized using inverse Fourier transforms and the overlap-add method.

To compute the binary weightings of the mask, the target and interfering signals were segmented into overlapping, windowed frames in the same way as the mixture, and the FFT was computed. If the sentence belonged to the un condition, a mask filled with ones was generated. Otherwise, to generate the IBM the FFT magnitudes were combined into 22 channels between 187.5 and 7937.5 Hz using non-overlapping rectangular windows and the same cutoff frequencies as the CI sound processor model. The IBM was assigned a value of one if the target-to-interferer energy was greater than or equal to a threshold (and zero otherwise), where the threshold was chosen to be equivalent to the SNR of the mixture signal. In other words, the masks were generated with a “relative criterion” of 0 dB (i.e., the difference between the threshold and mixture SNR was zero) so that the same IBM was obtained regardless of the SNR of the mixture (Kjems et al., 2009). If the sentence belonged to any of the conditions with mask errors, a model-generated mask (MGM) was generated from the IBM with the assigned levels of individual error types and T-F error structure using the method described in the next section. Finally, before applying the masks, they were expanded in the spectral dimension from 22 channels to 720 points by first reversing the rectangular windowing that was applied to the frequency bins to compute the IBM. Then, the FFT bins before 187.5 Hz and after 7937.5 Hz were assigned the same value as that of the first and last mask channel, respectively, and the resulting gain values were mirrored to produce the symmetry required for the inverse transforms.

5.2.2 Binary masks

To generate a large corpus of example masks that have errors with roughly the same statistical properties as authentic estimated masks, errors were artificially added to the IBMs using the model described in Chapter 4. Using the model to create masks as opposed to using authentic masks allows for evaluation of the consequences of different error properties in a controlled way. For example, with the model it is possible to generate a mask with exactly 10% false positive errors and 20% false negative errors, and then create another mask from the same IBM with exactly 20% false positive errors and 10% false negative errors. Tuning an authentic estimation algorithm to generate two masks for the same sentence with these precise error properties is largely impractical because algorithm parameters do not usually map directly and independently to the individual error properties of interest.

The basic mechanism of the model for this study can be described as follows. A new artificial mask is initialized by making a copy of the IBM, and then the individual T-F units are modified to produce a mask with the desired error rates and structure. This is accomplished by deciding for each individual T-F unit whether or not to flip its label by considering the probability of making that error based on the false positive or false negative rate as well as by considering how likely it is to have the same label of the units that are its direct neighbors.

The MGMs in this study were generated using the same methods as in Chapter 4. The masks contained one of sixteen different error configurations with false positive rates ranging from 0% to 40%, false negative rates ranging from 0% to 40%, and amounts of clustering at $\gamma = 1.0$ (i.e., unstructured errors) to $\gamma = 2.0$ (i.e., the amount of clustering that is typically introduced by existing IBM estimation algorithms; Chapter 4). There were four configurations with only false positives ($\alpha = 20\%$ and 40% with $\gamma = 1.0$ and 2.0), four configurations with only false negatives ($\beta = 20\%$ and 40% with $\gamma = 1.0$ and 2.0), and eight configurations with both false positives and false negatives ($\alpha = 20\%$ and 40% each combined with $\beta = 20\%$ and 40% at both $\gamma = 1.0$ and 2.0). Taken together with the IBM and the un, there were eighteen conditions in total. For demonstration, two example model-generated masks are shown in Figs. 19a-19c. Note that the errors in the MGM with $\gamma = 2.0$ are substantially more clustered in time and frequency than the errors in the MGM with $\gamma = 1.0$.

5.2.3 Participants

Eight CI recipients using Cochlear Limited devices with CI24RE implants or later participated in this study. All participants were at least 18 years of age, were native Australian English speakers, had at least six months of experience with an implant device, and were paid for volunteering. The age of the participants ranged from 32 to 84 years with an average age of 62 years. Implant use on the tested ear ranged from 0.9 to 8.7 years with an average of 5.9 years, and all but one participant was

post-lingually deafened. Table 2 outlines the biographical data for the participants.

5.2.4 Setup

The study was conducted using a computer system that emulates a basic version of each participant’s sound processor with their own personal fitting settings. The system included a performance real-time “target” machine from Speedgoat™, which was responsible for executing the real-time model of a sound processor, and a “host” computer, which was used to program the target computer and to run the testing software. The host computer interfaced with the target computer using the Mathworks™ Simulink and xPC target framework in a similar manner to the hearing aid system described by Buchholz (2013). The Simulink model mimicked the behavior of the Nucleus® 5 and 6 systems without the directional microphone technology, the automatic scene classifier, and the latest noise and wind reduction technologies (Mauger et al., 2014). Thus, the primary components included a spectral flattening filter, T-F decomposition, n -of- m channel selection using ACE (see Section 3.2), and loudness growth and current level mapping. The hardware required to connect the xPC system to the CI was purpose-built and provided by Cochlear Limited.

In a typical setup with this real-time system, instead of wearing their normal behind the ear (BTE) sound processor, participants wear a BTE sound processor that has been modified to contain only the microphone components. The signal from the microphone in the BTE shell is sent to the target computer, and the model on the target computer processes the signal and determines the output stimulation pattern. The stimulation pattern is then sent to a box where it is encoded as a radio frequency signal and subsequently delivered to the participant’s receiver through a coil that has been affixed to their head with their magnet. The receiver decodes the radio frequency signal and transmits the stimulus data to the participant’s electrode array.

However, the setup in this study was atypical in that the participants did not wear the BTE shell during testing. Instead, the input signal to the target computer was acquired in one of two ways: either through a lapel microphone worn by the researcher conducting the study or from a RME Fireface UC sound card on the host computer that relays the playback of audio signals from the testing software. The lapel microphone allowed the researcher to communicate with the participant during the session, but the microphone was silenced whenever the testing software playback was active to ensure that participants were not disturbed by environmental sounds during the test.

This preprocessing and playback setup was necessary because the MGMs could not be generated in real-time. To illustrate the relationship between the binary masks generated during the preprocessing with the T-F representation of energy after acoustically reconstructing the signal for playback, the spectrograms from within

Table 2: Biographical data for the cochlear implant recipients who participated in this study.

Participant	Age (yrs)	Implant age (yrs)	Implant type	Processor	Maxima	Stim rate (Hz)	Number of chan	Pulse width (μ s)	SRT (dB)
P1	40	6.5	CI24RE(ST)	CP900 Series	8	900	22	25	+2
P2	32	8.5	CI24RE(ST)	CP900 Series	8	900	22	25	+6
P3	72	4.7	CI512	CP900 Series	10	900	22	25	+2
P4	66	8.7	CI24RE(ST)	CP900 Series	8	900	22	25	+8
P5	56	0.9	CI422	CP900 Series	8	900	22	25	+6
P6	84	8.5	CI24RE(CA)	CP900 Series	12	900	20	25	+3
P7	73	1.6	CI422	CP810	8	900	22	37	+10
P8	75	7.5	CI24RE(ST)	CP900 Series	10	900	22	25	+4

the CI model during playback are shown in Figs. 19d-19f for the masks in Figs. 19a-19c. Although there was inevitably some mismatch between the intended binary mask and the resulting representation of the T-F energy in the CI model because of the acoustic reconstruction, the level of mismatch appears to be relatively small considering that the general properties of the mask errors (e.g., amount of error, locations of errors, and amount of structure in the errors) were seemingly well-maintained. For further demonstration of the setup, Figs. 19g-19i show the resulting stimulation sequences (i.e., electrodiagrams) from the CI model for each of these sentences.

5.2.5 Procedure

Before each of the participants arrived on the day of their session, the Simulink model was configured and validated to mimic the individual's everyday sound processor settings. All adaptive processing (including Adaptive Dynamic Range OptimisationTM, automatic gain control, beamformers, and noise reduction) was disabled. Upon arrival, the participants were briefed about the study, and the researchers obtained the participant's written consent. Participants were instructed to remove their own sound processor, and the coil from the real-time system was affixed to the participant's head. The microphone input level was adjusted until the participant felt that it was at a comfortable listening level.

To familiarize with the speech recognition task, one list was played in quiet to the participant via the sound card input, and the participant was instructed to repeat back what they heard. The researcher and the participant sat facing each other in a quiet room, so the participant responses were scored by morpheme (as defined by the BKB-like corpus) as they were spoken.

Next, two SRTs were measured using an adaptive procedure (Keidser et al., 2013) with four lists of the BKB-like corpus (two lists for each measurement) and the interferer babble signal continuously looping in the background to determine the SNR that resulted in the participant receiving 50% intelligibility. A beep was presented before each sentence at a level that was 10 dB higher than the noise to cue the participant that a sentence would be presented.

In the final phase of the session, fixed-level testing was conducted for the binary masking conditions under comparison using the set of preprocessed sentences with an SNR that was equivalent to the SNR of the participant's average SRT (Table 2). Note that because the sentences were preprocessed for this portion of the testing, the interfering noise was not continuously running in the background, and therefore, performance in the un condition at the SRT SNR was generally lower than the 50% intelligibility that was obtained during the measurement of the SRTs. Each list of sentences was played to the participant in the randomized order at a level that was calibrated to be equivalent to 65 dB SPL for participants P1-P7 and 60 dB SPL for participant P8 (an SRT was unattainable for this participant when the noise was at 65 dB SPL) while the performance was recorded. For consistency,

a beep was presented before each sentence 10 dB higher than the signal to cue the participant that a sentence would be presented. Breaks were taken between each list, and the entire session lasted between approximately two and four hours. All testing was conducted using computer software developed by the National Acoustic Laboratories.

5.3 Results

Fig. 20 shows the measured speech recognition scores for the individual participants and the mean speech recognition for all participants. Three-way ANOVA with repeated measures indicated a significant effect of the clustering parameter γ ($F [1.04, 7.25] = 36.43, p < 0.001$), a significant effect of the false negative rate β ($F [2.00, 14.00] = 32.49, p < 0.001$), but a non-significant effect of the false positive rate α ($F [1.20, 8.37] = 1.86, p = 0.211$). It indicated a significant effect of the interaction between γ and α ($F [2.08, 14.59] = 34.87, p < 0.001$), a significant effect of the interaction between γ and β ($F [1.41, 9.86] = 23.56, p < 0.001$), and a significant effect of the interaction between α and β ($F [1.53, 10.68] = 17.65, p = 0.001$). Finally, the analysis also indicated a significant effect of the interaction between all three factors ($F [1.39, 9.74] = 43.64, p < 0.001$). Because the result of the Mauchly's test of sphericity was significant for all but the effect of β , a Greenhouse-Geisser adjustment was made to the respective degrees of freedom.

Although there is a considerable amount of variability in the overall levels of each participant's performance, the basic trends within each participant demonstrate significant similarities. In almost all cases, speech recognition substantially decreased when the mask errors were clustered (i.e., $\gamma = 2.0$). Moreover, with the exception of the participant P7, false positive and negative errors were nearly always both detrimental to intelligibility. The statistical significance of all main effects and interactions except the effect of α corroborates that these relationships are unlikely to have occurred by chance alone. Further, by removing P7's scores from the statistical analysis, the effect of α becomes highly significant ($p = 0.004$ from $p = 0.211$) while the significance of all of the other main effects and interactions also remain highly significant.

The inconsistency of participant P7's scores with the other participants' trends may be due to the fact that P7's sentences were mixed at 10 dB SNR (versus 2-8 dB for the other participants) to match P7's high SRTs in the first phase of the session. Because the interfering babble was so low in comparison to the target signal as a result of the high SNR, false positives may not actually have had much of an impact on P7's resulting stimulation sequences since the energy in those T-F units may have been low enough to cause the channel selection strategy to prioritize stimulating other channels. Additionally, given that the P7's speech recognition scores in the un condition in the second phase of the session were much higher than P7's scores for all of the processed conditions as well as substantially higher than the 50% recognition P7 obtained in the first phase, P7 may have been intolerant

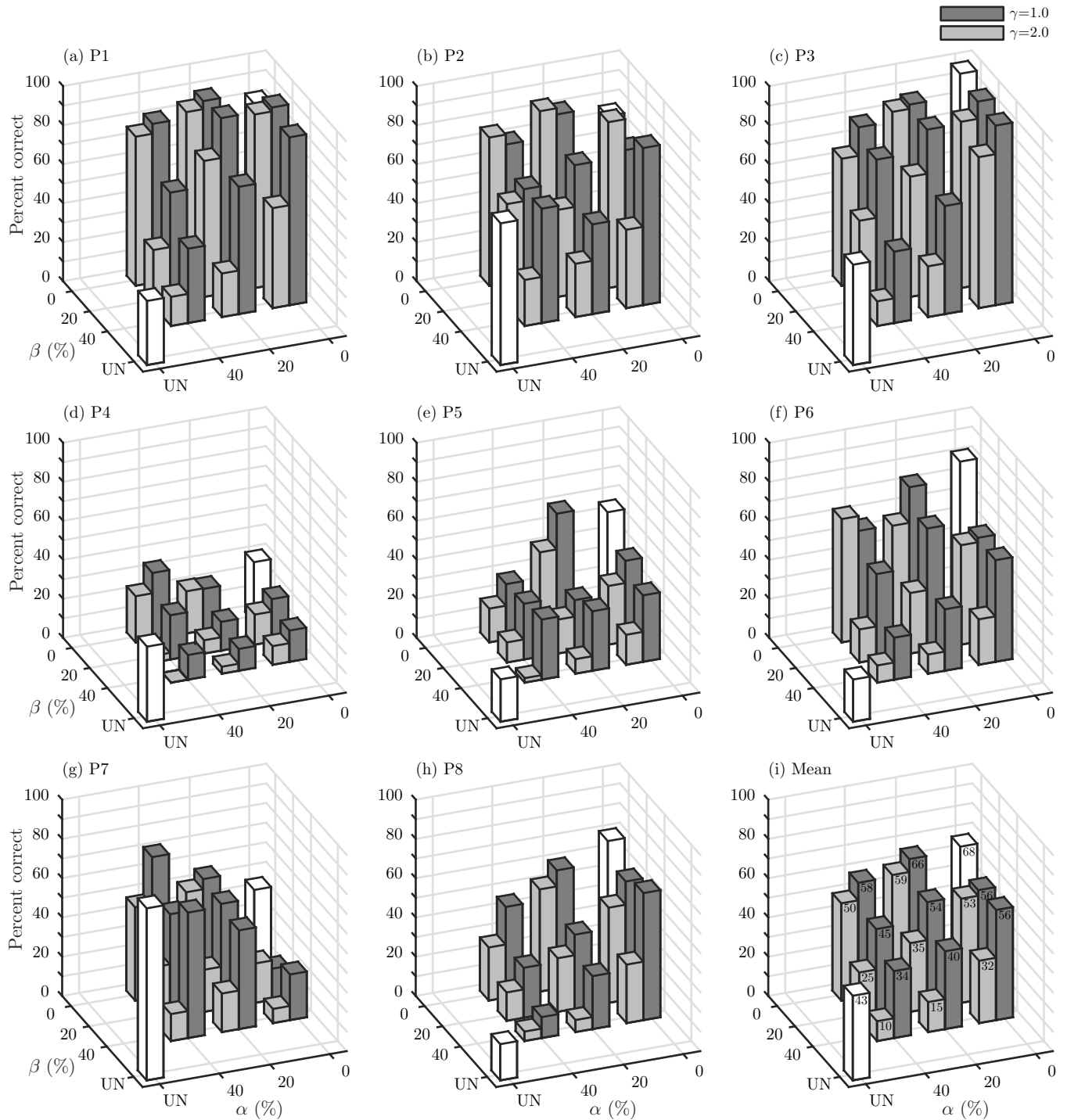


Figure 20: (a-h) Individual participant speech recognition scores and (i) mean speech recognition scores across all participants.

to T-F gain manipulation artifacts generally, subject to acclimatization during the test, or just inconsistent with this speech recognition task.

5.4 *Discussion*

Comparing the mean scores here for CI recipients with those of normal-hearing listeners in the three listener studies of Chapter 4, the effects of α , β , and γ on speech intelligibility, as well as their interaction, are remarkably alike. The first listener study in Chapter 4 investigated the effect of structure in false positive errors and showed that increased clustering resulted in increased sensitivity to false positive errors. This same tendency is demonstrated for CI recipients in the five conditions with $\beta = 0\%$. Next, the second listener study in Chapter 4 investigated the effect of structure in false negative errors. The speech recognition scores from their listener study not only replicated the plateau relationship between false negative errors and intelligibility that was first demonstrated in Li and Loizou (2008), but went on to show that the plateau relationship is only valid when the errors are unstructured because false negative errors can be as detrimental as false positive errors when they contain structure. In this study with CI recipients, the very same tendencies were demonstrated in the five conditions with $\alpha = 0\%$. Finally, the third listener study in Chapter 4 investigated the effect of interacting errors and structure. The speech recognition scores from their study established that the tolerance of false positive errors decreases when false negative errors are present and vice versa. In this study with CI recipients, speech recognition scores for a given false positive error rate also substantially decreased when false negative errors were present and vice versa. Taken together, there is substantial evidence showing that for both normal-hearing and CI listeners, clustering of binary mask errors reduces intelligibility outcomes; false negative errors can be just as detrimental to intelligibility as false positive errors when the errors contain structure or when both error types are present; and lastly, the individual tolerances of false positive and negative errors decrease when the other error type is present.

Although this study demonstrates the effect of error clustering as well as the relative effects of false positive and negative errors, the experimental setup was designed in a way that required acoustic reconstruction of the signals after applying the binary mask. In a more realistic scenario, the T-F gain manipulation would likely take place directly on the CI spectrograms within the sound processor, unless for example, the acoustic signals were preprocessed first with a hearing aid chip. However, the potential artifacts of the acoustic reconstruction such as musical noise or cross-talk between adjacent frequency channels most likely have limited impact on the overall conclusions of the study since the conclusions are drawn primarily from the differences between the conditions rather than on the actual values of the speech recognition scores. Moreover, because sounds lower than the threshold sound pressure level (by default, $T_{SPL} = 25$ dB) are truncated, the potential effects of any reconstruction artifacts were anyway reduced.

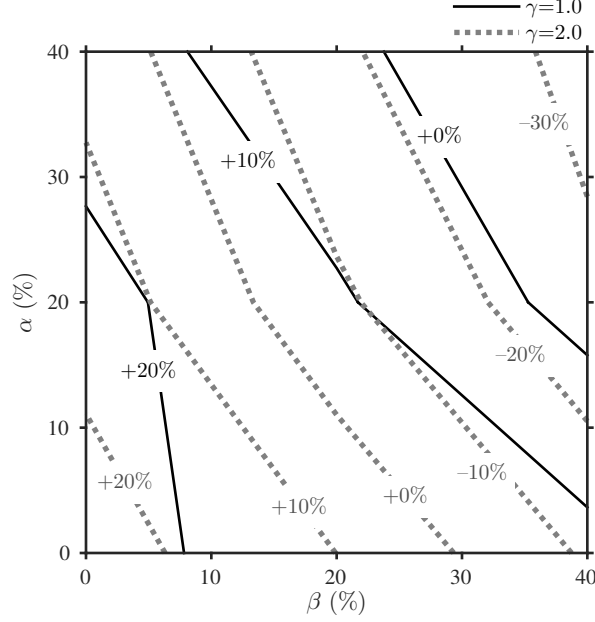


Figure 21: Contours of the mean change in speech recognition scores relative to the unity mask (UN) condition. When mask errors are unstructured ($\gamma = 1.0$), combinations of α and β on or below the solid contour lines indicate mean speech recognition changes equal to or better than the indicated amount. In contrast, when the mask errors are clustered with $\gamma = 2.0$, combinations of α and β on or below the dashed contour lines indicate mean speech recognition changes equal to or better than the indicated amount.

To deduce the implications of this intelligibility study on the classification accuracy requirements of IBM-based noise reduction algorithms, Fig. 21 shows a comparison of the mean speech recognition scores for combinations of α and β with unstructured ($\gamma = 1.0$) and structured ($\gamma = 2.0$) errors. According to this contour plot, all combinations of α and β below the solid line marked +0% would lead to an improvement in mean speech recognition scores if the errors are unstructured. On the contrary, if the errors are clustered in a way that makes it twice as likely for neighboring T-F units to be in the same state as opposed to different states, then only the combinations of α and β below the dotted line marked +0% would lead to an improvement in mean speech recognition scores. Although the data that makes up this contour plot is limited, it suggests that the requirements on binary mask estimation algorithms are more rigorous when the algorithm makes errors that are more clustered. To give an explicit example, a mask with a 20% false positive rate and a 20% false negative rate is likely to yield a +10% average speech recognition improvement if the errors are unstructured but a -10% change in average speech recognition scores if the errors are clustered.

Quantitatively speaking, a 20% false positive rate and 20% false negative rate of unstructured errors resulted in an estimated change in speech recognition of about

+11% for CI recipients, but an estimated change of about +41% for normal-hearing listeners (Chapter 4). Similarly, a 20% false positive rate and 20% false negative rate of clustered errors resulted in an estimated change in speech recognition of about -8% for CI recipients but an estimated change of about $+22\%$ for normal-hearing listeners. Thus, it appears that the CI listeners obtained lower intelligibility improvements than their normal-hearing counterparts for masks with a given accuracy rate, at least with respect to the testing methods and speech material used in the respective studies. In order for the CI recipients to obtain an improvement in word recognition, the masks with clustered errors needed to contain combinations of error rates that approximately satisfied $\alpha + \beta \leq (30\% - 45\%)$. Contrastingly, in order for the normal-hearing listeners to obtain an improvement in word recognition, the masks with clustered errors needed to contain combinations of error rates that approximately satisfied $\alpha + \beta \leq (55\% - 60\%)$. These results support previous conjectures that CI recipients are more sensitive than normal-hearing listeners to false positive or false negative errors (Qazi et al., 2013, Mauger et al., 2012).

From a channel selection perspective, the implications from this study are slightly less clear, largely because of the complex interaction between binary mask errors and the n -of- m channel selection strategy in the CI model. With the default sound processor settings for example, only eight channels are stimulated in each cycle. Thus, an electrodogram from a mask-processed sentence would have many more false negatives than the mask itself if the mask had time frames containing more than eight correctly labeled target-dominated T-F units, and the electrodogram would have far fewer false positives than the mask itself if the mask had time frames containing more than eight incorrectly labeled interferer-dominated T-F units. Furthermore, whenever there is enough signal energy to activate more than eight channels (which is generally true whenever speech is present), introducing a false positive inevitably introduces a false negative and vice versa. Therefore, in order to understand the effects of particular error rates as well as error clustering specifically in the framework of channel selection as opposed to noise reduction, further investigation is needed with an experimental setup that removes the constraints of an n -of- m strategy. Despite the complexity of this relationship between mask errors and structure with channel selection errors and structure, the speech recognition data of the CI recipients in this study were astoundingly similar to the speech recognition data of the normal-hearing listeners. Taken all together, it is quite clear that structure has a substantial impact on the intelligibility of speech processed or encoded with binary decisions.

To summarize, this study investigated the influence of both structured and unstructured binary mask errors on speech recognition scores in CI recipients. Clustering of the binary mask errors consistently reduced the listeners' ability to tolerate binary mask errors in binary-masked speech, and in general, the individual tolerances of each of the error types decreased when both types of errors were present.

Moreover, false negatives were as detrimental as false positives, and at least according to the means, false negatives were actually more detrimental than false positives when the errors were clustered. These trends all replicate those of normal-hearing listeners (Chapter 4). However, the CI recipients tended to be more sensitive than normal-hearing listeners to false positive or false negative errors, at least with respect to the testing methods and speech material in the respective studies.

All things considered, since clustering is a relatively superficial definition of structure in that it does not directly relate to the content in the underlying signals, it may not be appropriate to translate the error tolerances reported in this study directly to realistic estimation algorithms since realistic estimation algorithms likely contain other kinds of structure that effect tolerances in different ways. It is clear, however, that structure does substantially impact how much error is acceptable, and since realistic algorithms will make errors in structured ways, it is critical to consider the structure when analyzing potential outcomes. Going forward, it will be important to establish an understanding of how more tangible forms of structure impact intelligibility, such as determining how vital it is to maintain the integrity of the transitions between speech segments and gaps or how crucial it is to preserve the temporal and spectral shape of formants and transients. The approach presented in this work provides a foundation for these kinds of future studies, and as a result, provides a promising direction for developing more accurate objective analysis techniques and for developing more focused noise reduction and channel selection algorithms.

CHAPTER VI

OBJECTIVE ANALYSIS OF STRUCTURE IN ESTIMATED BINARY MASKS AND THE IMPLICATION ON INTELLIGIBILITY PREDICTIONS

Structure in the errors of binary masks artificially generated by a statistical model of binary masks substantially reduced the ability of the masks to improve speech intelligibility for both normal-hearing listeners and cochlear implant recipients. To investigate how this effect of structure impacts authentic estimated binary masks, this chapter objectively analyzes the error characteristics of binary masks generated by a wide range of estimation algorithms. Furthermore, this chapter also evaluates the effectiveness of prevalent binary masking objective measures, such as the metric based on the difference between the hit rate and the false alarm rate (i.e. H-FA), to predict the effects of the different error characteristics on intelligibility outcomes. Based on the analysis, it is evident that the artificially induced range of error characteristics among the model-generated masks is to likely exist to a similar degree among authentically-generated masks, and therefore, that the measured effect of structure on speech intelligibility using model-generated masks is likely to translate to real-world algorithms. As a result, it is clear that since H-FA, the commonly used metric for evaluating estimation algorithms to date, cannot predict the effect of structure on intelligibility, estimation algorithms should not be evaluated and compared solely with it.

6.1 *Introduction*

Chapters 4 and 5 demonstrated that underlying structure in mask errors can significantly influence intelligibility outcomes. To conduct this study, masks were artificially created using a statistical model that can generate masks from IBMs with specific amounts of false positives and negatives that are clustered together in varying amounts. These MGMs were designed based on the assumption that estimation algorithms make both error types at the same time and that they make errors that contain different amounts of clustering. However, it is unclear to what extent estimation algorithms actually differ in their error rates and underlying structure.

In a large portion of binary mask estimation literature, algorithms are compared solely on their average classification performance using the H-FA objective measure (see Section 2.5.1). This is largely due to the fact that Kim et al. (2009) reported a moderate correlation ($r = 0.80$) between H-FA and speech intelligibility in their listener study with one estimation algorithm. More recently however, Taal et al. (2011b) conducted a thorough evaluation of traditional objective measures for predicting the intelligibility of reconstructed binary masked speech and continued on to develop the STOI measure (Taal et al., 2011a) (see Section 2.5.2). In spite of the development of STOI, many researchers who are working to develop binary mask estimation algorithms continue to use H-FA for performance evaluation even with its limited validation. Taking into account the demonstration of the effect of structure on speech intelligibility outcomes in Chapters 4 and 5 and the failure of H-FA to properly guide the development of a successful sparsity-based estimation algorithm in Chapters 2 and 3, the effectiveness of analysis with H-FA has come into question, particularly because it does not take into account any structure in mask errors.

The main contributions of this chapter are to present an objective analysis of the error characteristics in a collection of estimated binary masks (EBMs) and to demonstrate the inability of H-FA to account for differences in underlying structure within mask errors. The investigation focuses first on exploring how well the range of error rates and clustering in the MGMs tested in the listener studies in Chapters 4 and 5 represent the distributions among authentic EBMs in order to permit a better understanding of the implications of the results from the listener studies. To accomplish the investigation, the statistical model for binary masks first introduced in Chapter 4 is employed to facilitate quantifying the amount of clustering within the masks from each algorithm. Additionally, the model is extended to be able to quantify clustering in the temporal and spectral directions independently in order to be able to further describe the underlying error characteristics of these algorithms. The second part of the investigation focuses on analyzing the effectiveness of H-FA and STOI to predict the intelligibility outcomes from Chapter 4.

6.2 Methods

The analysis was divided into two separate sections. In the first section, estimated masks from a collection of binary mask estimation algorithms were examined to assess the distribution of error characteristics in authentic masks. In the second section, prevalent objective measures were evaluated on their ability to predict the influence of these different error characteristics on speech intelligibility outcomes.

6.2.1 Analysis of masks

6.2.1.1 Binary mask estimation algorithms

The present study evaluated five different binary mask estimation algorithms that represent a range of proposed approaches (Table 3). Four of these algorithms employ a supervised-learning strategy, in which knowledge about the feature distribution of target- and interferer-dominated T-F units gained during training is used to classify T-F units encountered during testing. Specifically, the distribution of a set of normalized AMS features is learned by a GMM classifier (Kim et al., 2009, May and Dau, 2014a). These approaches will be referred to as *pre-trained*.

Three of these pre-trained algorithms use an updated set of AMS features that are computed with an auditory-inspired modulation filterbank (May and Dau, 2014a) in place of the linearly-scaled, constant-bandwidth modulation filterbank proposed first by Kim et al. (2009). Furthermore, one of these pre-trained algorithms is trained on a wide range of noise types (referred to as *general*) rather than employing separate classifiers for each type of background noise (referred to as *specific*). Finally, the last main difference between the pre-trained estimators is whether or not the strategies include a SVM-based integration stage, which considers the spectro-temporal context of neighboring T-F units. Such an integration has been shown to substantially improve the overall classification accuracy (May and Dau, 2014a). This integration stage is designed specifically to incorporate the information of nearby T-F units into the classification decision so that, for example, the probability of a single T-F unit being target-dominated is dependent on the probability of the neighboring units also being target-dominated.

In contrast to pre-trained strategies, a more generic strategy that does not rely on an initial training stage will be referred to as *untrained* in this context. In the untrained strategy included in this study, the probability of speech is estimated in individual T-F bins by continuously estimating the background noise power spectral density (Gerkmann and Hendriks, 2012). An estimate of the IBM is then obtained by applying a noise-specific threshold to the estimated speech presence probability (SPP) (May and Gerkmann, 2014). This algorithm has shown superior classification performance in comparison to pre-trained methods when the algorithms are tested with unseen noise types.

The GMM-based segregation systems were trained with 180 randomly selected sentences from the IEEE corpus mixed with seven different types of noise at three

Table 3: List of estimation algorithms and their primary differences.

Algorithm	Class	AMS filterbank	Noise treatment	T-F integration	Citation
A1	pre-trained	linearly-scaled	specific	no	May and Dau (2014a)
A2	pre-trained	logarithmically-scaled	specific	no	May and Dau (2014a)
A3	pre-trained	logarithmically-scaled	general	no	May and Dau (2014a)
A4	pre-trained	logarithmically-scaled	specific	yes	May and Dau (2014a)
A5	untrained	—	specific	no	May and Gerkmann (2014)
Key comparison	A2 vs. A5	A1 vs. A2	A2 vs. A3	A2 vs. A4	

different SNRs (-5 , 0 , and 5 dB SNR). The following seven maskers were used: stationary speech-shaped noise (ICRA1-noise) and speech-modulated noise (ICRA-7-noise) (Dreschler et al., 2001); 8 amplitude-modulated pink noise; traffic noise and classical music taken from B  chler (2002); and destroyer noise and non-stationary factory noise from Varga and Steeneken (1993). The threshold which separates the training material into target-dominated and interference-dominated T-F units on the basis of the prior SNR was set to -5 dB. The same training material was used to select the noise-specific thresholds for the SPP-based segregation system. The SVM-based integration stage consisted of a plus-shaped, causal spectro-temporal integration window spanning over three adjacent time frames and nine frequency channels (May and Dau, 2014a). A linear SVM classifier was trained on all background noises using a small set of 10 IEEE sentences mixed at -5 , 0 and 5 dB SNR (May and Dau, 2014a).

6.2.1.2 Model for quantifying structure

To evaluate the amount of structure in the estimated binary masks, the model first described in Chapter 4 was employed (see Section 4.2). In that first formulation of the model, the potential functions were parameterized by the set θ consisting of A , B , γ , and λ . A was used to parameterize the potential functions that control the bias towards mislabeling interferer-dominated units, and B was used to parameterize the potential functions that control the bias towards mislabeling target-dominated units. The parameter γ was used to parameterize the potential functions that control the amount of clustering between neighboring T-F units in both the temporal and spectral directions, and the parameter λ was used to parameterize the potential functions that control the bias towards being in the interferer-dominated state. Finally, to enable more intuitive probabilistic descriptions of the error rates in the model, the false positive rate α and the false negative rate β were used in place of the parameters A and B , respectively, to describe how often mislabeling occurs in particular masks. Of note though is that α is monotonically related to A , and β is monotonically related to B .

In this chapter, the parameters α and β are still used to describe the false positive and negative rates of non-ideal masks relative to the matching ideal mask. However, γ can now take on a new formulation in which the clustering is decoupled in the temporal and spectral directions. In this new formulation, the parameterization set θ consists of the parameters A , B , γ_t , γ_f , and λ . Thus, the probability of a particular assignment of the set of variables using the new model formulation can

be represented by the following joint distribution $p(\mathbf{x}, \mathbf{y})$,

$$\begin{aligned}
 p(\mathbf{x}, \mathbf{y}) = & \frac{1}{Z} \prod_i \phi_i(x_i) \prod_i \phi_i(y_i) \cdots \\
 & \prod_{\langle i,j \rangle_t} \phi_{\langle i,j \rangle_t}(x_i, x_j) \prod_{\langle i,j \rangle_f} \phi_{\langle i,j \rangle_f}(x_i, x_j) \cdots \\
 & \prod_{\{i,j\}_0} \phi_{\{i,j\}_0}(x_i, y_j) \prod_{\{i,j\}_1} \phi_{\{i,j\}_1}(x_i, y_j). \tag{11}
 \end{aligned}$$

The potential functions $\phi_i(x_i)$ and $\phi_i(y_i)$ give a non-negative weight to each possible value of the associated random variable, and the potential functions $\phi_{\langle i,j \rangle_t}(x_i, x_j)$, $\phi_{\langle i,j \rangle_f}(x_i, x_j)$, $\phi_{\{i,j\}_0}(x_i, y_j)$, and $\phi_{\{i,j\}_1}(x_i, y_j)$ give a non-negative weight to each possible combination of values between the pair of associated random variables. This is exactly the same as in the first formulation in Chapter 4 except that instead of using the potential function $\phi_{\langle i,j \rangle}(x_i, x_j)$ for all pairs of neighboring T-F units, $\phi_{\langle i,j \rangle_t}(x_i, x_j)$ is employed when the neighboring T-F units are connected in the temporal direction and $\phi_{\langle i,j \rangle_f}(x_i, x_j)$ is employed when the neighboring T-F units are connected in the spectral direction. For demonstration, Fig. 22 shows example MGMs with (b) symmetrical unstructured errors (i.e., $\gamma = 1.0$ or equivalently, $\gamma_t = 1.0$ and $\gamma_f = 1.0$), (c) symmetrically clustered errors (i.e., $\gamma = 2.0$ or equivalently, $\gamma_t = 2.0$ and $\gamma_f = 2.0$), (d) errors clustered primarily in the temporal direction ($\gamma_t = 4.0$ and $\gamma_f = 1.0$), and (e) errors clustered primarily in the spectral direction ($\gamma_t = 1.0$ and $\gamma_f = 4.0$). The collection of parameters and potential functions are summarized in Table 4 for reference.

To do parameter estimation with the model, a graphical model was configured for a set of non-ideal masks and each T-F unit was assigned two binary indicator features to identify whether the target-dominated or interferer-dominated label was the correct label. Note that the masks were typically truncated beforehand to ensure that all masks within a dataset were of the same size. Next, all of the weights in each of the potential functions were tied to the value one or to one of the set of free weights depending on the type of potential function and its position within the mask. The potential functions were then computed by exponentiating the assigned weights, which ensures that the optimization program becomes convex.

The minimization of the regularized negative log likelihood was subsequently performed using the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (Liu and Nocedal, 1989) (see Section 4.3.1).

$$\arg \min_{\theta} \left(-\log p_{\theta}(\mathbf{x}|\mathbf{y}) + \Lambda \|\theta\|_2^2 \right) \tag{12}$$

During each iteration of the optimization program, the estimation of Z and the marginal probabilities were performed as a subroutine using the mean field method (see Section 4.3.1). The regularization parameter Λ was heuristically chosen to be

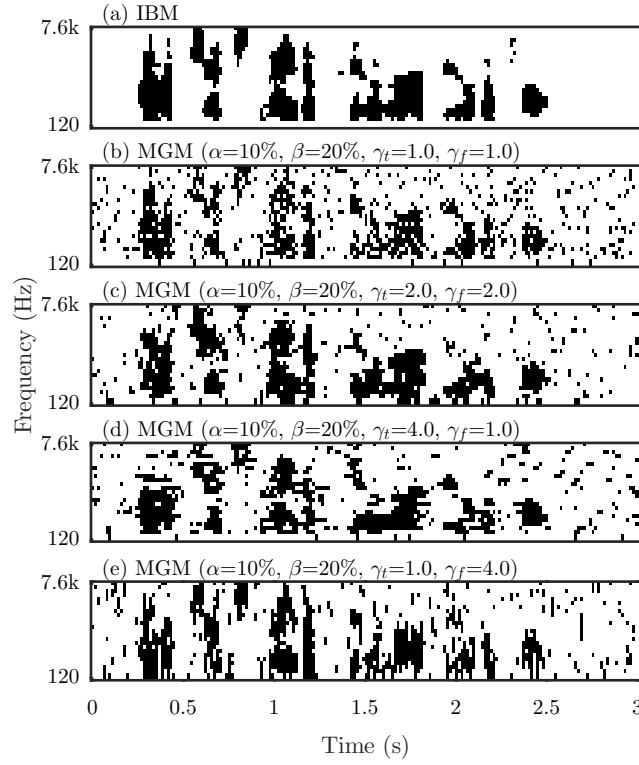


Figure 22: Example (a) ideal binary mask (IBM), (b) model-generated mask (MGM) with symmetrical unstructured errors, (c) MGM with symmetrically clustered errors, (d) MGM with errors clustered primarily in the temporal direction, and (e) MGM with errors clustered primarily in the spectral direction.

Table 4: List of model parameters.

Notation	Description
$\gamma, \gamma_t, \gamma_f$	Controls the amount of clustering between neighboring T-F units in both the temporal and spectral directions, just the temporal direction, or just the spectral direction, respectively; $\gamma \geq 1, \gamma_t \geq 1, \gamma_f \geq 1$
α	False positive rate; $0 \leq \alpha \leq 1$
β	False negative rate; $0 \leq \beta \leq 1$
A	Controls the bias towards mislabeling interferer-dominated units; $0 \leq A \leq 1$
B	Controls the bias towards mislabeling target-dominated units; $0 \leq B \leq 1$
λ	Controls the bias towards being in the interferer-dominated state; $0 \leq \lambda \leq 1$
θ	Describes the set of model parameters, which consists of A, B, λ , and γ or the set of γ_t and γ_f
x_i	T-F unit in the non-ideal mask; takes values $[0\ 1]$ for the interferer- and target-dominated labels, respectively
y_i	T-F unit in the ideal mask; takes values $[0\ 1]$ for the interferer- and target-dominated labels, respectively
$p(\mathbf{x}, \mathbf{y})$	Joint probability distribution for the model; see Eq. 11
Z	Normalizing constant that forces the probability distributions to one and ensures it is valid
$\phi_i(x_i), \phi_i(y_i)$	Gives a non-negative weight to each possible value of the T-F units in the non-ideal and ideal masks, respectively; $[\lambda\ 1-\lambda]$
$\phi_{\langle i,j \rangle_t}(x_i, x_j),$ $\phi_{\langle i,j \rangle_f}(x_i, x_j)$	Gives a non-negative weight to each possible combination of values for T-F units in the non-ideal mask neighboring in any direction, the temporal direction, or the spectral direction, respectively; $[\frac{\gamma}{1}\ \frac{1}{\gamma}], [\frac{\gamma_t}{1}\ \frac{1}{\gamma_t}], [\frac{\gamma_f}{1}\ \frac{1}{\gamma_f}]$
$\phi_{\{i,j\}_0}(x_i, y_j),$ $\phi_{\{i,j\}_1}(x_i, y_j)$	Gives a non-negative weight to each possible combination of values for the T-F unit pairs in the same T-F position when the ideal T-F unit is interferer-dominated or target-dominated, respectively; $[1-A\ A], [B\ 1-B]$

10. Including this regularization term resulted in more stable computations during optimization. Nevertheless, the same approximate mean values for the weights were obtained with and without the regularization. At the completion of the optimization, the amount of clustering is, by definition, the value of the edge potential between x_i nodes that are in the same state relative to the value of the edge potential between x_i nodes that are in different states. When the model is configured for symmetrical clustering, this is γ . When the model is configured to decouple the clustering, γ_t will be defined by the edge potentials between T-F units that neighbor in the temporal direction, and γ_f will be defined by the edge potentials between T-F units that neighbor in the spectral direction.

6.2.2 Analysis of metrics

In order to evaluate the objective measures most well-known for the evaluation of binary mask estimation algorithms, the objective measures were assessed on their ability to predict the intelligibility scores from the listener studies in Chapters 4. In these listener experiments, speech mixed with babble was processed with MGMs containing varying levels of error rates α and β between 0% and 40% and the clustering parameter γ equal to 1.0 and 2.0 (i.e., γ_t was always equal to γ_f in the study). Ten normal-hearing listeners were scored on their word recognition of the processed signals. For the evaluation of the objective measures in this chapter, the masks and processed signals were regenerated using the same procedures.

6.3 Results

6.3.1 Analysis of masks

The investigation of the influence of more realistic mask errors in Chapters 4 and 5 was based on two basic assumptions. The first assumption was that estimation algorithms tend to make both false positive and negative errors at the same time and also that algorithms would do so at varying rates. The second assumption was that estimation algorithms tend to make errors that contain structure, and also presumably that algorithms would do so in varying ways and in varying amounts. The first part of this analysis addresses the validity of this first assumption, and the second part of this analysis addresses the validity of this second assumption.

6.3.1.1 Distribution of error rates

Fig. 23 depicts the spread of α and β rates over all of the estimated masks. The mean α and β combination for each algorithm is shown by the location of the algorithm abbreviation in the scatter plot, and the marginal histograms of the rates for each are shown along the axes. First and foremost, this illustration reveals that estimation algorithms do make both false positive and negative errors simultaneously and that they do so at vastly different rates.

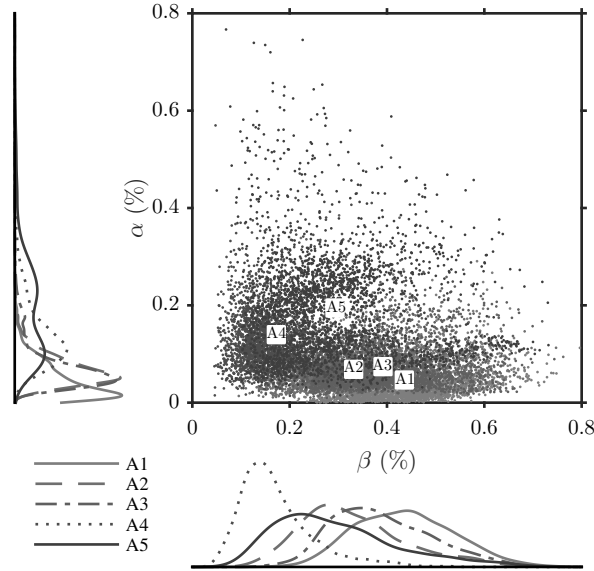


Figure 23: Two-dimensional scatter plots with marginal histograms of the false positive rate (α) and false negative rate (β) of each estimated binary mask, grouped by algorithm. The algorithm labels on the scatter plot are placed on the mean α and β for the algorithm.

Furthermore, Fig. 23 also reveals that false negative rates are generally much higher than the false positive rates among this collection of algorithms. This is especially true for the more traditional approaches A1, A2, and A3. On the other hand, A4 and A5 tend to have a more equal balance of false positive and negative error rates. Taken together, majority of the masks contain false negative rates ranging from about 5% to 65% and false positive rates ranging from about 0% to 40%, which actually exceeds the range of error rates investigated in Chapters 4 and 5.

The marginal distributions of false negative rates are exceptionally broad for nearly all of the algorithms meaning that although masks from the A2 algorithm, for example, on average have 30% or 35% false negative rates, 50% or more of the target-dominated regions in many of these masks are incorrectly removed. With regard to the specific differences between the algorithms, the untrained algorithm A5 tends to contain more false positives than the pre-trained algorithms A1-A4, but generally less false negative errors than the more basic pre-trained algorithms (A1-A3). Using the logarithmically-scaled AMS filterbank reduces the amount of false negative errors (A2 versus A1), and classifiers trained on specific noise types compared to many noise types also reduces the amount of false negative errors (A2 versus A3). The T-F integration both improves the mean and dramatically reduces the spread of false negative rates but also slightly increases the mean and the spread of the false positive rates (A4 versus A2).

In the literature to date, these types of estimation algorithms have primarily

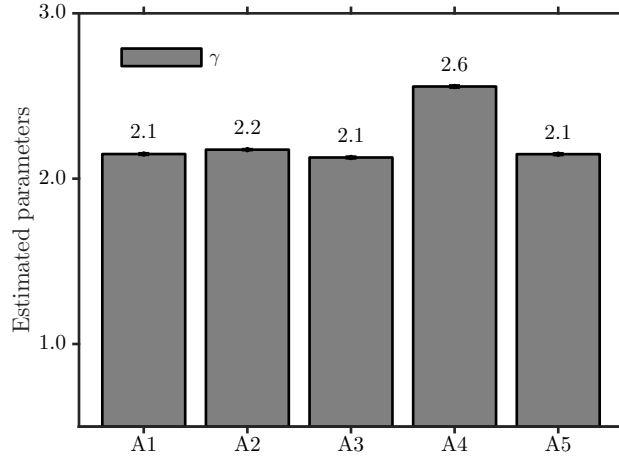


Figure 24: Mean estimated γ (plus and minus the standard error of the mean) over randomly selected subsets of the estimated binary masks (EBMs).

been evaluated by the mean false positive and negative rates without accounting for the distribution of the rates or, probably more importantly, without accounting for where in the speech signals the errors are made or how one error in the mask is related to the other errors in the mask. Since Chapters 4 and 5 have demonstrated that the strength of the conditional dependence between neighboring T-F units in a mask can substantially influence how many errors listeners can tolerate in binary-masked speech, the next section is dedicated to analyzing how strong the conditional dependence is among these estimated masks, as well as how much this dependence changes between algorithms.

6.3.1.2 Distribution of clustering

For each EBM algorithm, 100 masks were randomly selected and the amount of clustering (γ) in the set of masks was estimated. This process was repeated 20 times. Fig. 24 shows a bar plot of the mean estimated γ for each of the algorithms, as well as the standard errors of the means.

Among this collection of estimation algorithms, γ was relatively stable around 2.1 to 2.2 no matter which class, AMS filterbank, or noise treatment. This aligns well with the amount reported in Chapter 4. However, when the algorithm included a T-F integration stage, γ increased to 2.6, which is a change in magnitude of approximately 0.5. Thus, although the T-F integration tended to reduce the amount of errors in the mask, it simultaneously increased the amount of clustering among the mask errors. Given the substantial differences in the ability of listeners to tolerate errors when γ changed from 1.5 to 2.0 or from 2.0 to 2.5 in the studies in Chapter 4, an increase of γ from 2.2 to 2.6 is noteworthy. Having said that, from Fig. 23 it is clear that this integration stage markedly reduced the amount of false

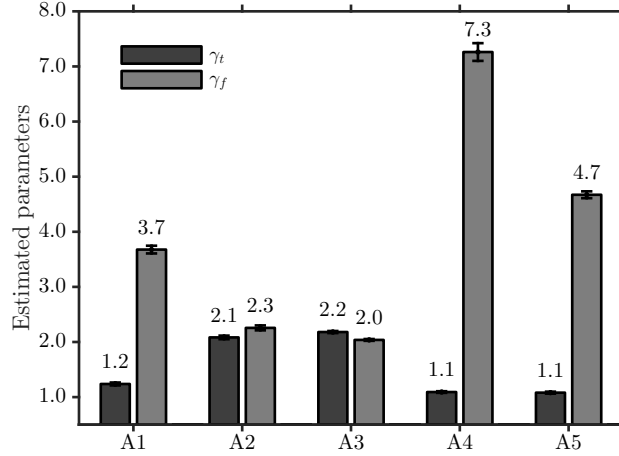


Figure 25: Mean estimated γ_t and γ_f (plus and minus the standard error of the mean) over randomly selected subsets of the estimated binary masks (EBMs).

negative errors, and it concurrently increased the amount of false positive errors, but to a smaller degree. Of note is that, based on the studies in Chapters 4 and 5, the tolerance of false positive errors appears to be less impacted by clustering than false negative errors so this shift may be less detrimental than it seems at first look.

This evaluation procedure was repeated again to estimate the amount of clustering separately in the temporal (γ_t) and spectral (γ_f) directions. Fig. 25 shows a bar plot of the mean estimated γ_t and γ_f for each of the algorithms, as well as the standard errors of the means.

The estimated values of γ_t and γ_f differ drastically in range from those obtained for γ . Because this new formulation of the model differs from that used in the listener studies in Chapters 4 and 5, it is unclear what the implications are for differences with these magnitudes. However, it is evident that different estimation approaches not only produce different *amounts* of clustering in their errors, but they also produce errors with clustering tendencies that have substantially different orientations. Most noteworthy of this difference is in the algorithm with a T-F integration stage (A4) where clustering in the spectral direction substantially exceeds clustering in the temporal direction. However, the untrained algorithm A5 also tends to cluster primarily in the spectral direction and, although to a lesser degree, A1 does as well. On the other hand, the specific and general classifiers that use the logarithmically-scaled AMS filterbank (A2 and A3) tend to make errors that are clustered relatively equally in the temporal and spectral directions.

The perceptual effects of spectrally orientated clustering compared to symmetrically oriented clustering is currently unknown. However, it has been shown that increased clustering of mask errors symmetrically reduces the number of errors that both normal-hearing listeners and CI recipients can tolerate while still being able to understand the target speech (Chapters 4 and 5). Since the analysis in this section

has plainly demonstrated that authentic estimation algorithms contain critically different amounts and types of clustering, it is crucial for binary masking objective measures to account for these kinds of differences. In the next section, two objective measures are evaluated specifically on their capacity to predict the influence of symmetrically oriented clustering since this is one of the only studies to date reporting speech intelligibility scores for binary masks with a range of underlying structure in the mask errors.

6.3.2 Analysis of metrics

Fig. 26a illustrates a contour plot of the behavioral word recognition scores for speech processed with MGMs from the third listener study in Chapter 4. For each combination of α and β , the plot gives an estimate of the mean word recognition score depending on if the mask contains unstructured errors ($\gamma = 1.0$) or errors that are clustered with $\gamma = 2.0$. In Chapter 4, this same contour plot was shown in mean word recognition scores relative to the scores for the un condition. However, because H-FA cannot predict intelligibility scores for the un condition, the scores in this chapter are instead reported in their absolute forms.

Based on this contour plot, if the errors in the MGMs are unstructured ($\gamma = 1.0$), all combinations of false positive rates and false negative rates that fall on or below the solid contour line marked, for example, 50% would lead to mean word recognition scores of 50% or better. In contrast, if the errors in the MGMs are clustered with $\gamma = 2.0$, only combinations of false positive rates and false negative rates that fall on or below the dashed contour line marked 50% would lead to mean word recognition scores of 50% or better. Taken together, this data suggests that, on average, masks with higher amounts of clustering must achieve higher accuracy rates in order to yield similar intelligibility outcomes. Furthermore, because the slopes of the solid contour lines in Fig. 26a are nearly equal to -1 , masks containing unstructured errors generally appear to be equally influenced by false positive and negative errors. On the other hand, the dashed contour lines are more steeply sloping. This suggests that higher false negative error rates are more detrimental to intelligibility outcomes than false positive error rates when the errors are clustered.

To put this data into the perspective of the estimation algorithms analyzed in the previous section, algorithms A1, A2, A3, and A5 (all estimated to contain errors with $\gamma \approx 2.0$) would need to achieve classification accuracy rates such that $\alpha + \beta \leq 20\%$ in order to yield average word recognition scores of 50% or better (which is an approximate improvement of 30% compared to the un condition). However, the only algorithm which comes close to achieving this level of accuracy is A4, which is estimated to contain errors with a higher amount of clustering and therefore, may actually require even higher levels of accuracy to yield 50% or better word recognition scores.

Figs. 26b and 26c plot contours of the predicted intelligibility outcomes using H-FA and STOI, respectively, for the regenerated stimuli. The general qualitative

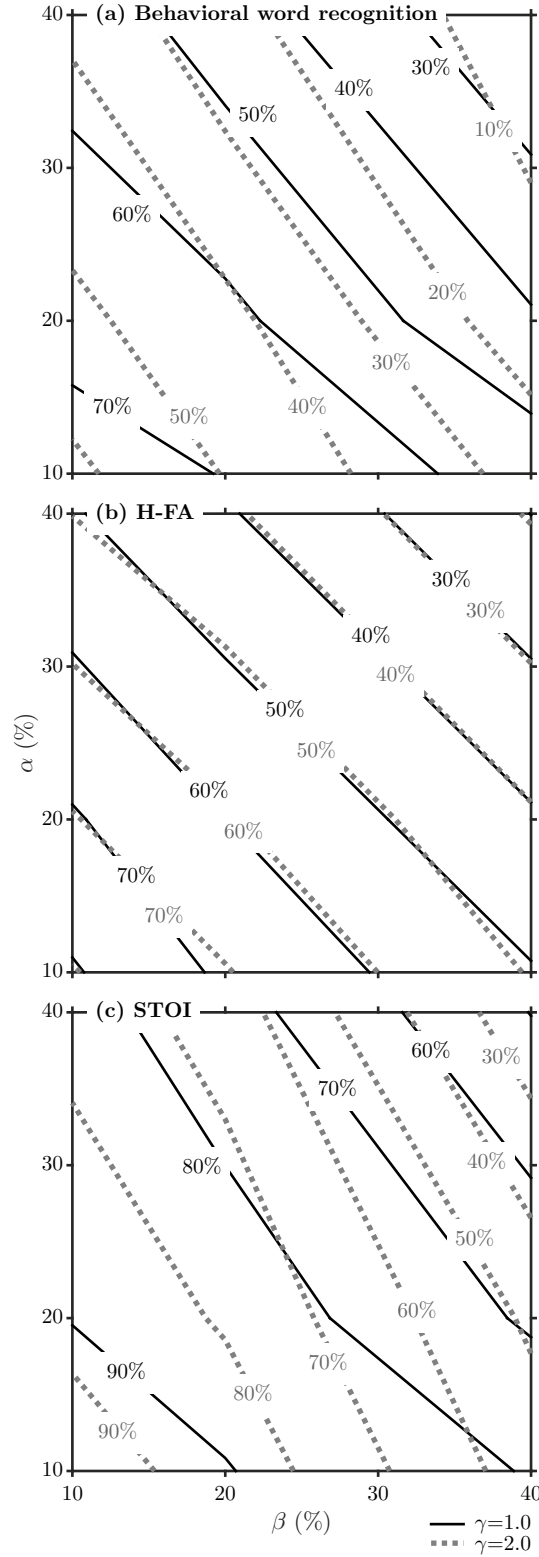


Figure 26: Contours of (a) behavioral word recognition from Chapter 4, (b) predicted intelligibility scores using H-FA, and (c) predicted intelligibility scores using STOI for speech processed with model-generated masks (MGMs) that contain a range of false positive rates (α), false negative rates (β), and two levels of clustering (γ). Masks with unstructured errors ($\gamma = 1.0$) are indicated with solid contour lines, whereas masks with clustered errors ($\gamma = 2.0$) are indicated with dashed contour lines.

relationship between false positive rates, false negatives rates, and intelligibility outcomes are predicted very well for the conditions with unstructured errors by both H-FA and STOI, as depicted particularly by the fact that the slopes of the solid contour lines all have approximate slopes of -1 . However, H-FA fails to predict the negative impact that the clustering of the errors has on the intelligibility of the binary-masked speech, as demonstrated by the lack of increased requirements on the accuracy of the masks with $\gamma = 2.0$ as well as the lack of increased steepness in the contour lines for the masks with $\gamma = 2.0$. This result is not surprising given that H-FA does not take into account any of the structure in the masks; it simply uses classification accuracy to predict intelligibility. Although this approach appears to predict outcomes relatively accurately among masks with unstructured errors or even among masks with the same kind of structure, it fails to predict differences that arise in intelligibility due to higher level differences in the masks.

STOI, on the other hand, does quite well to predict the negative impact of clustering, at least qualitatively, as shown by the fact that the contour lines for the masks with $\gamma = 2.0$ are much lower in magnitude than the nearby contour lines for the unstructured masks. STOI also successfully predicts the stronger influence of false negative errors on intelligibility in comparison to false positive errors, as evidenced by the more steeply sloping dashed lines.

To recap, STOI predicts intelligibility by taking into account the correlation of the envelopes between the reference and degraded signals over time segments of about 400 ms. When random T-F units are incorrectly labeled, the envelope of the degraded signal is not notably altered; in contrast, when many T-F units temporally aligned are incorrectly labeled, the envelope of the degraded signal is severely impacted so that the correlation with the reference signal is much lower. When many short-time segments have significantly degraded envelopes, the predicted intelligibility outcome is appropriately low.

Of note however, is the fact that STOI does not specifically account for spectrally oriented clustering since it does not evaluate correlation across frequency explicitly. Since the results in Fig. 25 suggest that some estimation algorithms tend to make errors that cluster primarily in the spectral direction, this may be particularly important. Until the perceptual effects of the orientation of clustering are more clearly understood however, the importance of this omission by STOI will remain unsettled.

6.4 Discussion

The analysis presented in this chapter suggests that estimation algorithms can produce masks with errors that tend to be at substantially different rates and tend to cluster in substantially different ways. When taken in the context of the listener studies with MGMs containing a range of error rates and a range of clustering with similar spans (Chapters 4 and 5), it is evident that the detrimental effect of clustering on listeners' ability to tolerate errors is not an artificial phenomena that only

exists in the laboratory. On the contrary, the measured effect of clustering on intelligibility using artificial masks is relevant to authentic masks in the real world because the real-world algorithms actually produce errors with considerably different amounts of clustering. Presumably then, the algorithms that produce errors with large amounts of clustering must achieve lower rates of false positives and negatives than algorithms with lower amounts of clustering in order to yield the same intelligibility.

Given this affirmation that clustering has implications in real-world algorithms, it was a logical next step to follow with an investigation into what happens when the clustering parameter is decoupled in the temporal and spectral directions. Based on the results presented in Fig 25, it is evident that errors can be clustered both equally and unequally in the temporal and spectral directions, depending on the type of estimation algorithm. However, it remains to be seen how making errors that are clustered primarily in the spectral direction affects intelligibility in comparison to making errors that are clustered equally in both directions.

Nonetheless, clustering is a superficial definition of structure in that it does not relate directly to the content in the underlying signals. Fig. 27 looks at a concrete, yet still simple form of structure: the distribution over the SNRs in each T-F unit. Using the A2 algorithm as an example, Fig. 27a shows a histogram of the SNRs of T-F units with false positive and negative errors as well as a histogram of the SNRs from all of the T-F units. This figure illustrates that the A2 algorithm tends to make more errors in T-F units that have an SNR near the IBM threshold since the distribution for the errors is more heavily weighted toward the threshold in comparison to the distribution for all of the T-F units. Hence, it suggests that the likelihood of a false positive or negative error is highly dependent on the proximity of the SNR of the T-F unit to the IBM threshold.

Fig. 27c shows the same histograms for a set of MGMs generated with the same amount of γ clustering as the A2 masks. Although the clustering parameter γ primarily introduces clustering among the mask errors, it also introduces the tendency to make more errors near the transition between target- and interferer-dominated regions. In this way, γ implicitly captures the relationship between errors and local SNR that is observed in the A2 masks. For comparison, Fig. 27e shows the histograms for another set of MGMs generated with the estimated γ_t and γ_f for A2, and Fig. 27g shows the histograms for a third set of MGMs generated with $\gamma = 1.0$. Despite the additional parameter for the masks in Fig. 27e, the distribution over SNRs for this set of masks matches almost exactly those of the A2 masks in Fig. 27c with a single γ parameter. However, both of the MGM sets appear to be less kurtotic (i.e., “peaky”) than the A2 masks themselves, particularly with respect to the distribution of the false negative errors. Thus, it seems that the MGMs are making more errors in T-F units with really high SNRs than the A2 estimated masks do. Nevertheless, as expected the MGMs that were generated with $\gamma = 1.0$ have a distribution of errors that matches that of the underlying distribution. Since the

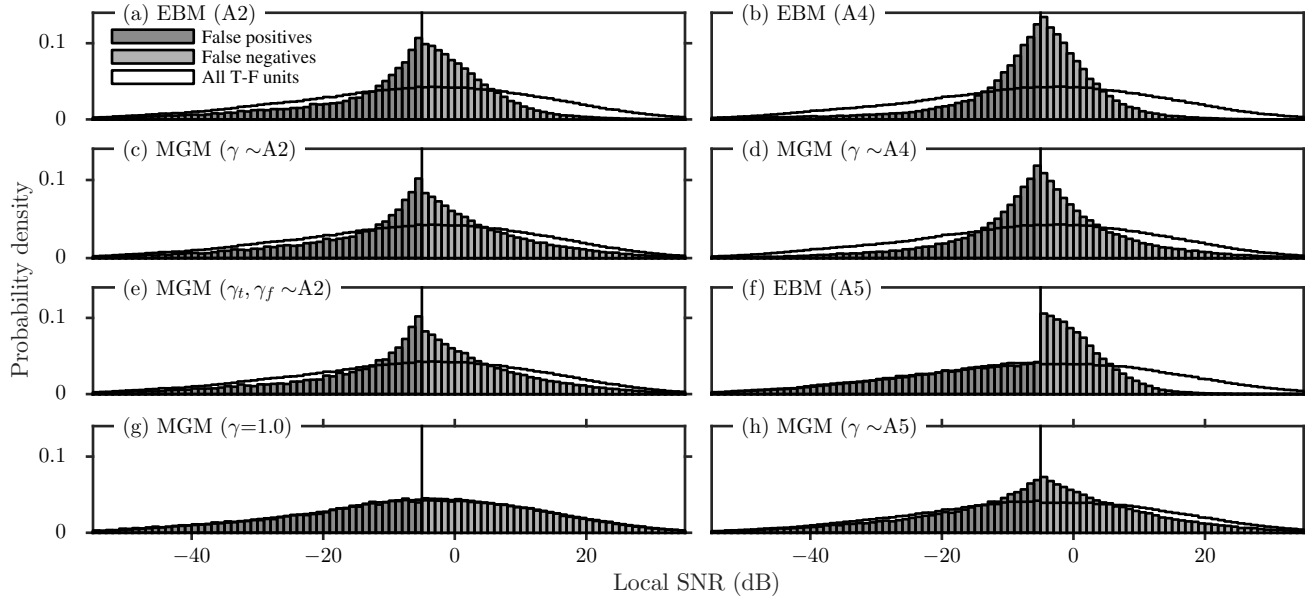


Figure 27: Histograms of the local SNRs from the T-F units with false positive and false negative errors in a set of (a) estimated masks, (c) model-generated masks (MGMs) with γ equal to the estimated γ in the masks, (e) MGMs with γ_t and γ_f equal to the estimated γ_t and γ_f in the masks, and (g) MGMs with $\gamma = 1.0$. For reference, histograms of the local SNRs from all of the T-F units in the masks are shown underneath. (b) contains estimated masks, (d) model-generated masks (MGMs) with γ equal to the estimated γ in the masks, (f) estimated masks, (c) model-generated masks (MGMs) with γ equal to the estimated γ in the masks.

distribution of the errors over local SNR for the MGMs with $\gamma = 1.0$ is the same as the underlying distribution of local SNR over all T-F units, the masks generated with $\gamma = 1.0$ contain errors that are indeed unstructured.

Figs. 27b and 27f show the histograms for the set of A4 estimated masks with T-F integration and untrained A5 estimated masks, respectively. Comparing the distribution for the A4 algorithm with the distribution for the A2 algorithm, the A4 estimated masks contain substantially less errors at the extreme SNRs. The set of MGMs generated with $\gamma \sim \text{A4}$ mimic this distribution very closely (Fig. 27d). On the other hand, the distribution for the A5 algorithm suggests that the false positives errors this algorithm tends to make are actually unstructured whereas the false negative errors are skewed towards the threshold in a similar way to the A2 algorithm. The set of MGMs generated with $\gamma \sim \text{A5}$ fail to capture these differences in distributions (Fig. 27h). Note that the set of MGMs generated with $\gamma_t, \gamma_f \sim \text{A5}$ also failed to capture this behavior since its distributions matched those in Fig. 27h almost exactly (not shown).

It is clear from this analysis that although it is a simple definition of structure, clustering can indirectly capture more tangible definitions of structure as well. Specifically, it was demonstrated that at least some estimation algorithms tend to make more errors in T-F units with SNRs near the threshold between the target- and interferer-dominated regions, and furthermore, that MGMs with $\gamma > 1.0$ also tend to make more errors in T-F units with SNRs near the threshold. Nonetheless, it was also demonstrated that clustering of errors fails to capture some higher levels of structure that are likely to influence intelligibility outcomes in different ways.

Broadly speaking, it is evident from this study that the MGMs in Chapters 4 and 5 showcase differences among the artificially-generated masks that likely exist to a similar degree in authentically-generated masks, and as a result, it is of considerable importance that H-FA, the commonly used metric for evaluating estimation algorithms, cannot predict the differences in intelligibility outcomes for masks that contain errors with differing levels of clustering. Thus, it is well substantiated to say that estimation algorithms should not be evaluated and compared with H-FA alone.

At least based on the analysis thus far, STOI seems to be a suitable replacement in situations when listener studies are impractical. However, it remains to be seen how spectrally oriented clustering impacts speech intelligibility outcomes, and since STOI is not currently designed to incorporate structural information across frequency channels in the same way that it does across time, it would not take into account this kind of structure. Beyond clustering and cluster orientation, it remains to be seen how more advanced definitions of structure (such as the distribution of error rates over SNR) impact speech intelligibility outcomes and whether or not STOI is able to capture them. Furthermore, despite STOI being a promising metric for binary masking evaluation, it does require acoustic reconstruction, which is particularly ill-suited for evaluation of binary masks in the context of cochlear

implants. It would be advantageous to develop a counterpart metric that accounts for structure in mask errors, but that does not require acoustic reconstruction.

This study demonstrated the realism of the masks that were artificially generated in Chapters 4 and 5 by comparing the artificially generated range of error characteristics with those of estimated masks from a collection of current estimation algorithms. This analysis substantiated that the conclusions from the listener studies in Chapters 4 and 5 about the influence of structure in binary mask estimation translates to authentic estimated binary masks. In addition to that, it highlighted some higher levels of structural variability among estimation algorithms, including the orientation of clustering and the conditional dependence of error rates on local SNR. The impact of these kinds of structure on speech intelligibility remain unclear.

Because the criteria on error rates is substantially lower when errors are unstructured, one possible interpretation to this study could be to conclude that algorithms should be designed so that they make unstructured errors. However, it is both improbable and impractical to do this, in part because the structure in the mask is likely to be what actually contains the information. Thus, the main conclusion should not be to design algorithms that tend to make errors with less structure, but that algorithms should be designed and assessed in more thorough ways than by simply comparing the accuracy of the classification of individual T-F units. Moreover, researchers are likely being too optimistic about the current performance of many state-of-the-art estimation algorithms since performance thus far has largely been evaluated with the H-FA metric without regard to the underlying structure in the mask or the underlying structure in the speech itself.

CHAPTER VII

CONCLUSION

This thesis presented a series of studies that primarily demonstrated the influence of structure in binary mask estimation error on speech intelligibility for both normal-hearing listeners and CI recipients. To start, Chapters 2 and 3 demonstrated the need for a better understanding of how binary mask errors impact speech intelligibility outcomes of binary-masked speech in the context of both noise reduction and CI channel selection. Subsequently, Chapter 4 introduced a novel approach for characterizing the effects of binary mask errors that employs concepts from the well-studied Ising model for image de-noising. Chapters 4 and 5 used this model to systematically study the influence of structure in binary mask errors by running intelligibility experiments with normal-hearing listeners and CI recipients, respectively. The listener experiments collectively demonstrated that structure in binary mask errors can substantially influence how many errors listeners can tolerate, and furthermore, that CI recipients may be more susceptible in general to these errors. Finally, Chapter 6 employed the binary mask model in another way to assess the structure in the mask errors that current estimation algorithms tend to make. In the end, this chapter provided an assessment of the impact of the listener studies on the development and evaluation of current and future approaches to T-F binary masking.

Broadly speaking, the collection of work in Chapters 2 through 6 not only introduced a novel and broad technique, but additionally showed that the technique is a valuable approach. The main contributions of this thesis with respect to the specific aims defined at the outset were addressed as follows:

Develop binary mask algorithms that leverage sparsity: Chapter 2 presented a new binary mask estimation approach that uses MP to identify the target-dominated T-F regions of speech mixed with interfering noise. The performance of the new estimation approach was evaluated with prevalent binary mask objective measures. Despite satisfying the criteria implied by Li and Loizou (2008), the speech intelligibility outcomes of the proposed algorithm were predicted to be low. As a result, the need for more specific and reliable design criteria was uncovered. Moreover, the notion that errors containing structure may impact intelligibility differently than unstructured errors was conceived. Subsequently, Chapter 3 presented an evaluation of a CI channel selection algorithm that also uses MP to identify the target-dominated T-F regions of speech mixed with interfering noise. To initially evaluate this strategy, the CI processor was modeled by speech vocoded with the

stimulation patterns and the resulting intelligibility of the vocoded speech was measured with an objective measure. Based on this analysis, the MP coding strategy was predicted to substantially improve speech recognition in noise in comparison to traditional channel selection strategies. To test this prediction, the speech recognition of normal-hearing listeners with the vocoded speech was measured in a listener study. Unexpectedly, the speech recognition outcomes with the proposed strategy were actually lower than the speech recognition outcomes with the traditional strategy. Again, the need for more specific and reliable design criteria was discussed, but this time with respect to CI strategies specifically.

Develop investigation framework for binary masks: As Chapters 2 and 3 discussed, practical binary mask estimation and channel selection strategies make errors that are not well-represented by the unstructured errors that were previously employed by other researchers to analyze the effect of binary errors on speech recognition. However, generating masks with specific amounts of each error type and specific kinds of structure is nontrivial. For example, tuning an authentic estimation algorithm to generate two masks for the same sentence with precise error properties is largely impractical because estimation algorithm parameters do not usually map directly and independently to the individual error properties of interest. In order to address this, Chapter 4 presented a novel investigation framework for artificially generating masks with specific error profiles and then demonstrated how to use this framework to generate artificial masks (i.e., *sampling*) and to characterize errors in a collection of masks (i.e., *parameter estimation*). Following this, the model was extended in Chapter 6 to highlight differences in higher levels of structure in the mask errors that authentic estimation algorithms tend to make.

Demonstrate influence of structure in binary masking errors: To start, Chapter 2 employed the binary mask model to evaluate the influence of individual error types and error structure in a systematic way. The results clearly illustrate the detrimental effect of structure on the tolerance of errors for normal-hearing listeners. Following this study, Chapter 5 extended the investigation of binary mask errors to establish the influence of both structured and unstructured errors on speech recognition in CI recipients. The trends in the results very closely follow those of the normal-hearing listeners and again, clearly demonstrate that underlying structure in mask errors significantly impacts the intelligibility outcomes of binary-masked speech. Moreover, the results suggest that CI recipients may be more susceptible to binary mask estimation error. Finally, Chapter 6 analyzed the implications of these new insights on the development and evaluation of algorithms that estimate the ideal mask. This study first showed that the masks artificially generated for the listener studies collectively represent a range of error profiles that also exist among authentically estimated masks. This validation of realism substantiates that the negative impact of structured errors has implications to the design and evaluation of estimation algorithms. Additionally, the effectiveness of the widely used objective

measure H-FA was discredited. Generally speaking, these studies jointly establish that noise reduction and channel selection strategies should be designed and assessed in more thorough ways than by simply comparing the accuracy of the classification of individual T-F units. It is clear that structure substantially impacts how much error is acceptable, and since realistic algorithms will make errors in structured ways, it is important to consider the structure when analyzing potential outcomes.

7.1 *Limitations*

Error clustering is a relatively superficial definition of structure. By focusing on this definition of structure, these studies have blatantly ignored more prominent notions of speech structure such as formants, formant transitions, and transients. It would be advantageous to be able to identify which of these more concrete definitions of structure are most influential, and as a result, be able to provide more insight about which tangible forms of information are most important to capture and when in the binary masking framework. Nonetheless, by demonstrating that even superficial definitions of structure can impact the tolerance of errors, the studies in this thesis jointly establish that noise reduction and channel selection strategies based on the binary masking approach should be designed and assessed in more thorough ways than is standard in literature to date.

7.2 *Future directions*

The sum total of this work provides new insights into some of the fundamental research questions that need to be addressed before future development of noise reduction and channel selection strategies based on the binary masking approach can be properly guided.

- First and foremost, the impact of the orientation of error clustering is currently unknown. Based on the fact that authentic estimation algorithms introduce errors that are orientated in substantially different directions, it will be worthwhile to conduct listener studies to discern the individual influences of temporal and spectral clustering.
- Depending on the outcome of such a study, a more effective objective measure should be designed, either by modifying STOI or using an entirely different approach, so that binary mask estimation algorithms can be more accurately evaluated before or in conjunction with listener studies.
- Building upon the discussion of the limitations of this thesis, it will be constructive to consider the influence of other more prominent and tangible forms of structure. Doing so will likely provide more functional knowledge about what makes the IBM actually successful in improving intelligibility rather

than simply providing theoretical information about how accurately the IBM needs to be replicated.

- Given that both of the listener studies in this thesis were conducted in the context of noise reduction, it would be valuable to assess the influence of binary mask errors specifically in the context of channel selection to uncover any differences in the speech intelligibility outcomes between binary masking as noise reduction and binary masking as channel selection. Since the concept of binary masking is actually very well-suited to CI processing in general and since the sparse coding model is also very well-suited to CI channel selection, this line of research likely contains the most potential for future innovation.

This body of future work will provide significantly more direction to the development of noise reduction and channel selection strategies by providing information about which binary decisions matter most and when. Moreover, there will be more specific guidance about how to most effectively tune classification to ensure high enough accuracy. Doing so may require taking into account higher order statistical structure in speech, perhaps beyond sparsity, to effectively distinguish the target speech from interfering signals in order to successfully improve speech understanding in noise.

References

- Mahnaz Ahmadi, Vauna L. Gross, and Donal G. Sinex. Perceptual learning for speech in noise after application of binary time-frequency masks. *The Journal of the Acoustical Society of America*, 133(3):1687–1692, 2013.
- Michael C Anzalone, Lauren Calandruccio, Karen A Doherty, and Laurel H Carney. Determination of the potential benefit of time-frequency gain manipulation. *Ear and Hearing*, 27(5):480–492, 2006.
- Aurele Balavoine, Justin Romberg, and Christopher J Rozell. Convergence and rate analysis of neural networks for sparse approximation. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(9):1377–1389, 2012.
- Richard G Baraniuk, Emmanuel Candes, Michael Elad, and Yi Ma. Applications of sparse representation and compressive sensing [scanning the issue]. *Proceedings of the IEEE*, 98(6):906–909, 2010.
- John Bench, Åse Kowal, and John Bamford. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British journal of audiology*, 13(3):108–112, 1979.
- Christopher M Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.
- Jesper B Boldt. *Binary Masking & Speech Intelligibility*. Dissertation, Aalborg University, 2011.
- Inge Brons, Rolph Houben, and Wouter A. Dreschler. Perceptual effects of noise reduction by time-frequency masking of noisy speech. *The Journal of the Acoustical Society of America*, 132(4):2690–2699, 2012.
- Douglas S Brungart, Peter S Chang, Brian D Simpson, and DeLiang Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120(6):4007–4018, 2006.
- Jorg M Buchholz. A real-time hearing-aid research platform (HARP): realization, calibration, and evaluation. *Acta Acustica united with Acustica*, 99(3):477–492, 2013.
- Michael Christoph Büchler. *Algorithms for sound classification in hearing instruments*. PhD thesis, Swiss Federal Institute of Technology, Zurich, Switzerland, 2002.

- Emmanuel J Candes and Michael B Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.
- Adam S Charles, Pierre Garrigues, and Christopher J Rozell. A common network architecture efficiently implements a variety of sparsity-based inference problems. *Neural computation*, 24(12):3317–3339, 2012.
- Fei Chen and Philipos C Loizou. Predicting the intelligibility of vocoded speech. *Ear and Hearing*, 32(3):331–338, 2011.
- Michael F Dorman, Philipos C Loizou, Jeanette Fitzke, and Zhemin Tu. The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *The Journal of the Acoustical Society of America*, 104:3583–3585, 1998.
- Wouter A. Dreschler, Hans Verschuure, Carl Ludvigsen, and Søren Westermann. ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *Audiology*, 40(3):148–157, 2001.
- J Garofolo, L Lamel, W Fisher, J Fiscus, D Pallett, N Dahlgren, and V Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1990.
- Timo Gerkmann and Richard C Hendriks. Unbiased mmse-based noise power estimation with low complexity and low tracking delay. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(4):1383–1393, 2012.
- Ray L Goldsworthy and Julie E Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6):3679–3689, 2004.
- Kun Han and DeLiang Wang. A classification based approach to speech segregation. *The Journal of the Acoustical Society of America*, 132(5):3475–3483, 2012.
- Eric W. Healy, Sarah E. Yoho, Yuxuan Wang, and DeLiang Wang. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 134(4):3029–3038, 2013.
- Hans-Günter Hirsch and David Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.
- Volker Hohmann. Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica United with Acustica*, 88(3):433–442, 2002.
- Yi Hu and Philipos C Loizou. A comparative intelligibility study of single-microphone noise reduction algorithms. *The Journal of the Acoustical Society of America*, 122(3):1777–1786, 2007.

- Yi Hu and Philipos C Loizou. A new sound coding strategy for suppressing noise in cochlear implants. *The Journal of the Acoustical Society of America*, 124:498–509, 2008.
- Yi Hu and Philipos C. Loizou. Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. *The Journal of the Acoustical Society of America*, 127(6):3689–3695, 2010.
- Gitte Keidser, Harvey Dillon, Jorge Mejia, and Cong-Van Nguyen. An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function. *International journal of audiology*, 52(11):795–800, 2013.
- Mead C Killion, Patricia A Niquette, Gail I Gudmundsen, Lawrence J Revit, and Shilpi Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(4):2395–2405, 2004.
- Gibak Kim and Philipos C Loizou. Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints. *Signal Processing Letters, IEEE*, 17(12):1010–1013, 2010a.
- Gibak Kim and Philipos C Loizou. Improving speech intelligibility in noise using environment-optimized algorithms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2080–2090, 2010b.
- Gibak Kim, Yang Lu, Yi Hu, and Philipos C Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3):1486–1494, 2009.
- Ulrik Kjems, Jesper B Boldt, Michael S Pedersen, Thomas Lunner, and DeLiang Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, 126(3):1415–1426, 2009.
- Ning Li and Philipos C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America*, 123(3):1673–1682, 2008.
- Leonid M Litvak, Anthony J Spahr, Aniket A Saoji, and Gene Y Fridman. Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners. *The Journal of the Acoustical Society of America*, 122:982, 2007.
- Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

- Philipos C Loizou and Gibak Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):47–56, 2011.
- Stephane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.
- Stefan J Mauger, Pam W Dawson, and Adam A Hersbach. Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction. *The Journal of the Acoustical Society of America*, 131(1):327–336, 2012.
- Stefan J Mauger, Chris D Warren, Michelle R Knight, Michael Goorevich, and Esti Nel. Clinical evaluation of the nucleus® 6 cochlear implant system: Performance improvements with smartsound iq. *International journal of audiology*, 53(8):564–576, 2014.
- Tobias May and Torsten Dau. Environment-aware ideal binary mask estimation using monaural cues. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4, 2013.
- Tobias May and Torsten Dau. Computational speech segregation based on an auditory-inspired modulation analysis. *The Journal of the Acoustical Society of America*, 136(6):3350–3359, 2014a.
- Tobias May and Torsten Dau. Requirements for the evaluation of computational speech segregation systems. *The Journal of the Acoustical Society of America*, 136(6):EL398–EL404, 2014b.
- Tobias May and Timo Gerkmann. Generalization of supervised learning for binary mask estimation. In *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, pages 154–158. IEEE, 2014.
- Vivienne L Ming and Lori L Holt. Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, 126(3):1312–1320, 2009.
- Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
- Obaid ur Rehman Qazi, Bas van Dijk, Marc Moonen, and Jan Wouters. Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility. *Hearing Research*, 2013.
- Nicoleta Roman and John Woodruff. Intelligibility of reverberant noisy speech with ideal binary masking. *The Journal of the Acoustical Society of America*, 130(4):2153–2161, 2011.

- Nicoleta Roman, DeLiang Wang, and Guy J. Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.
- EH Rothauser, WD Chapman, N Guttman, KS Nordby, HR Silbiger, GE Urbanek, and M Weinstock. IEEE recommended practice for speech quality measurements. *Audio and Electroacoustics, IEEE Transactions on*, 17(3):225–246, 1969.
- Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20(10):2526–2563, 2008.
- Samuel Shapero, Adam S Charles, Christopher J Rozell, and Paul Hasler. Low power sparse approximation on reconfigurable analog hardware. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 2(3):530–541, 2012.
- Samuel Shapero, Christopher Rozell, and Paul Hasler. Configurable hardware integrate and fire neurons for sparse approximation. *Neural Networks*, 45:134–143, 2013.
- Margaret W Skinner, Laura K Holden, Lesley A Whitford, Kerrie L Plant, Colleen Psarros, and Timothy A Holden. Speech recognition with the nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults. *Ear and Hearing*, 23(3):207–223, 2002.
- Evan Smith and Michael S Lewicki. Efficient coding of time-relative structure using spikes. *Neural Comput.*, 17(1):19–45, January 2005.
- Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(23):978–982, 2006.
- Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2125–2136, 2011a.
- Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. *The Journal of the Acoustical Society of America*, 130(5):3013–3027, 2011b.
- Cees H Taal, Richard C Hendriks, and Richard Heusdens. Matching pursuit for channel selection in cochlear implants based on an intelligibility metric. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 504–508, 2012.

- Steven van de Par, Armin Kohlrausch, Richard Heusdens, Jesper Jensen, and Søren Holdt Jensen. A perceptual model for sinusoidal audio coding based on spectral integration. *EURASIP J. Appl. Signal Process.*, 2005:1292–1304, 2005.
- Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.
- DeLiang Wang. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, pages 181–197. Springer, 2005.
- DeLiang Wang. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification*, 12(4):332–353, 2008.
- DeLiang Wang, Ulrik Kjems, Michael S Pedersen, Jesper B Boldt, and Thomas Lunner. Speech perception of noise with binary gains. *The Journal of the Acoustical Society of America*, 124(4):2303–2307, 2008.
- DeLiang Wang, Ulrik Kjems, Michael S Pedersen, Jesper B Boldt, and Thomas Lunner. Speech intelligibility in background noise with ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, 125:2336, 2009.
- Yuxuan Wang and DeLiang Wang. Cocktail party processing via structured prediction. In *Advances in Neural Information Processing Systems 25*, pages 224–232, 2012.
- Yuxuan Wang and Deliang Wang. Towards scaling up classification-based speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(7):1381–1390, 2013.
- Blake S Wilson and Michael F Dorman. Cochlear implants: a remarkable past and a brilliant future. *Hearing Research*, 242(1):3–21, 2008.

VITA

Journal publications

Abigail A Kressner, Tobias May, and Christopher J Rozell. Objective analysis of binary masks for predicting the intelligibility of binary-masked speech. *The Journal of the Acoustical Society of America*, in preparation. (**Chapter 6**)

Abigail A Kressner, Adam Westermann, Christopher J Rozell, and Jörg M Buchholz. Cochlear implant speech intelligibility outcomes with binary mask errors. *The Journal of the Acoustical Society of America*, in preparation. (**Chapter 5**)

Abigail A Kressner and Christopher J Rozell. Structure in time-frequency binary masking errors and its impact on speech intelligibility. *The Journal of the Acoustical Society of America*, accepted March 2015. (**Chapter 4**)

Abigail A Kressner, David V Anderson, and Christopher J Rozell. Evaluating the generalization of the hearing aid speech quality index (HASQI). *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(2):407415, 2013.

Conference publications

Abigail A Kressner and Christopher J Rozell. Speech understanding in noise provided by a simulated cochlear implant processor based on matching pursuit. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 14, 2013. (**Chapter 3**)

Abigail A Kressner, David V Anderson, and Christopher J Rozell. Causal binary mask estimation for speech enhancement using sparsity constraints. *Proceedings of Meetings on Acoustics*, 19:055037, 2013. (**Chapter 2**)

Abigail A Kressner, David V Anderson, and Christopher J Rozell. A novel binary mask estimator based on sparse approximation. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7497–7501, 2013. (**Chapter 2**)

Abigail A Kressner, David V Anderson, and Christopher J Rozell. Robustness of the hearing aid speech quality index (HASQI). In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 209212, 2011.

Adam S Charles, Abigail A Kressner, and Christopher J Rozell. A causal locally competitive algorithm for the sparse decomposition of audio signals. In *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE)*, pages 265-270, 2011.

Conference abstracts

Abigail A Kressner and Christopher J Rozell. The influence of structure in binary mask estimation error on speech intelligibility. In *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, CA, 2014. Selected for oral presentation.

Abigail A Kressner and Christopher J Rozell. Speech separation using matching pursuit for time-frequency masking. In *Signal Processing with Adaptive Sparse Structured Representations (SPARS) Workshop*, Lausanne, Switzerland, 2013.

Abigail A Kressner, Adam S Charles, and Christopher J Rozell. Causal locally competitive algorithm for the sparse decomposition of audio signals. In *Communications and Signal Processing, 2012 IEEE Womens Workshop on*, Banff, Canada, 2012.

Abigail A Kressner, David V Anderson, and Christopher J Rozell. Computational auditory models validate the intelligibility benefit of efficient filters. In *International Symposium on Auditory and Audiological Research (ISAAR)*, Nyborg, Denmark, 2011.

Abigail A Kressner, Christopher J Rozell, and David V Anderson. Predicting speech quality using a computational auditory model. In *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, CA, 2010.

Other reports

Abigail A Kressner. Auditory models for evaluating algorithms. Masters thesis, Georgia Institute of Technology, Atlanta, Georgia, USA, 2011.