A NOVEL BINARY MASK ESTIMATOR BASED ON SPARSE APPROXIMATION

Abigail A. Kressner, David V. Anderson, and Christopher J. Rozell

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, Georgia 30332 USA
{abbiekre,anderson,crozell}@gatech.edu

ABSTRACT

While most single-channel noise reduction algorithms fail to improve speech intelligibility, the ideal binary mask (IBM) has demonstrated substantial intelligibility improvements. However, this approach exploits oracle knowledge. The main objective of this paper is to introduce a novel binary mask estimator based on a simple sparse approximation algorithm. Our approach does not require oracle knowledge and instead uses knowledge of speech structure.

Index Terms— Ideal binary mask, sparse approximation, time-frequency masking, noise reduction, intelligibility

1. INTRODUCTION

State-of-the-art single-channel noise suppression algorithms do not improve the intelligibility of speech signals corrupted by noise. However, one algorithm in particular has shown significant improvements in intelligibility for normal- and impaired-hearing listeners—the ideal binary mask (IBM) [1, 2]. The IBM exploits oracle knowledge of the target and interferer signals to preserve only the time-frequency (T-F) regions that are target-dominated. Although the necessity of oracle knowledge makes the IBM an impractical algorithm for nearly all real applications, the significant increase in intelligibility makes the IBM a desirable benchmark for T-F masking algorithms trying to restore the intelligibility of noisy speech. While the IBM has been studied extensively in the literature (e.g., [3, 4, 5, 6, 7, 8, 9, 10]), there are yet few practical algorithms that can be used for its estimation in real-world applications.

By preserving only T-F regions that are target-dominated, the IBM creates a T-F signal representation that is more sparse (i.e., has fewer non-zeros) than the original mixture. Many recent advances in signal processing have revolved around the notion of sparsity, and numerous researchers in the signal processing community are developing methods to solve the sparse coding problem efficiently (i.e., in real-time and with low-power). Sparse coding has been used to enhance corrupted speech in a few recent studies [11, 12, 13, 14, 15]. However, to the authors' knowledge, no one has employed sparse approximation methods to directly estimate a T-F mask. The main objective of this paper is to introduce a novel binary mask estimator based on a simple sparse approximation algorithm.

This research was made with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

2. BACKGROUND

Sparse coding models have led to many state-of-the-art results in both the signal processing and computational neuroscience communities. These models treat a signal as a linear combination of elements from a (potentially over-complete) dictionary with the intent of finding an approximation using as few of the dictionary elements as possible. An audio signal, $x(t)$, is represented by a linear superposition of a basic set of kernels, $\phi_i(t), \ldots, \phi_M(t)$, which can be positioned arbitrarily and independently in time. The convolutional form of this model is given as

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

where $\tau_{m,i}$ 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 speech sounds efficiently, one needs to find the optimal set of $\phi_m(t)$ (*learning*), as well as find the optimal set of $s_{m,i}$ and $\tau_{m,i}$ (*encoding* or *inference*). With regard to learning, the optimal dictionary for efficiently representing speech signals is a family of gammatone-like functions [16]. With regard to inference, Matching Pursuit (MP) is a greedy algorithm designed to minimize the number of non-zero coefficients (i.e., the number of non-zero $s_{m,i}$) such that the reconstruction error is small [17]. Specifically for the convolutional model, MP will first choose the time-shifted basis that has the largest inner product with the signal, then subtract the contribution due to that time-shifted basis, and repeat the process iteratively until the signal is satisfactorily decomposed.

3. BINARY MASK ESTIMATORS

3.1. Ideal binary mask

The IBM is a relatively straight-forward algorithm. The general idea is to create a binary mask, which is defined in the T-F domain as a matrix of binary gain values. The gain is applied to the T-F representation of the mixture of target and interferer signals before recombination in a synthesis filterbank. To compute the binary mask, separate T-F representations of the target and interferer signals are obtained using either a short-time Fourier transform or a gammatone filterbank. For each T-F unit, the power levels of the target and interferer levels are computed to determine the local signal-to-noise ratio (SNR). T-F units with a local SNR above a pre-defined threshold are assigned a value of one in the mask and zero otherwise (Fig. 1b).

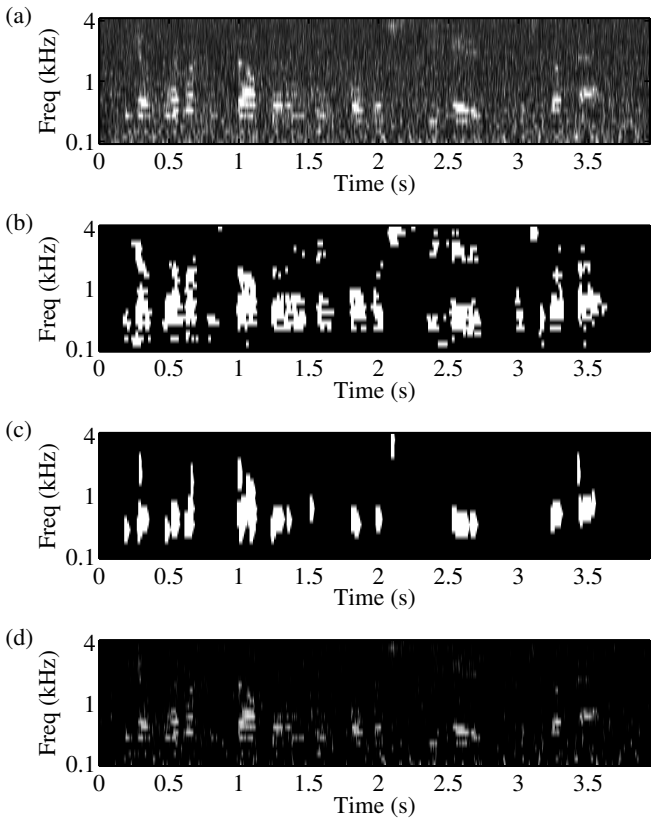

Fig. 1. Time-frequency representations of (a) a mixture (clean speech corrupted by pink noise at -5 dB SNR), (b) the ideal binary mask (0 dB threshold), (c) the MP binary mask (threshold of about 1.3), and (d) the FT binary mask (threshold of about 0.05). White indicates a value of one, and black indicates a value of zero.

3.2. Matching Pursuit binary mask

To compute the MP binary mask, we use MP and a dictionary consisting of gammatones to obtain a sparse approximation of the mixture. Since speech is efficiently encoded with gammatones, initial iterations in MP will likely choose coefficients that approximate speech energy rather than noise. Thus, by choosing a suitable stopping criteria for MP, the resulting set of coefficients will largely contain gammatones that fall in the T-F regions of the target speech. Instead of using the coefficients to synthesize a new signal (which may be impractical in certain applications), we use the sparse set of coefficients to identify target-dominated T-F regions and construct a binary mask (as described in the following paragraph). With this binary mask, we can modify the original signal rather than synthesize a new one.

Each coefficient MP chooses corresponds to a time-shifted gammatone. Since gammatones are localized to specific regions of frequency and time, we conclude that the regions corresponding to the chosen time-shifted gammatones likely contain target speech. Therefore, we set the binary mask for the union of these regions to one (Fig. 1c).

3.3. Filter-threshold binary mask

MP is highly nonlinear and requires a nontrivial level of computation and processing in order to account for structure in the signals. Given these drawbacks, we compare against a third mask estimator to look at what there is to gain by accounting for additional underlying structure. This estimator is loosely based on the filter and threshold (FT) algorithm [18]. FT is a very simple, causal approach to efficient audio coding 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. For our FT-based mask estimator, we compute filter response magnitudes for the mixture signal, and then simply set the mask to one if the response magnitude is greater than or equal to the threshold (Fig. 1d).

4. METHODS

Speech samples were created using the TIMIT speech corpus testing set re-sampled to 8 kHz [19]. A male- and female-spoken sentence was chosen from each of the eight dialect regions to form a set of sixteen sentences. To create the noisy signals, we added pink noise or realistic noise from the AURORA database (babble, car, street, and train) at input SNR levels of 0 dB and -5 dB.

We performed noise suppression on each of the stimuli using the three binary mask approaches described in Section 3. All algorithms used the same gammatone filterbank (24 4th-order filters spaced one ERB apart between 100Hz and 4 kHz, each with one-ERB bandwidth) [20]. Analysis was performed in the T-F domain, and we applied masks point-wise to the filter response of the mixture signal. To do the reconstruction, each frequency band of the modified response was delayed and scaled such that the peaks of the impulse response of each band had a maximum at 4 ms [20]. All of the frequency bands were then added together to obtain a single waveform.

For the IBM, we computed the filter response magnitudes for the clean and noise signals, and then summed the energy in each band within 20 ms time frames (Hamming window with 50% overlap). For each band, we assigned the mask a value of one at all 160 time samples within the time frame if the target energy was greater than or equal to the interferer energy scaled by a threshold factor; note that we assigned T-F units a value of one if the criterion was met in either of the overlapping frames. We used a threshold of -Inf dB to simulate the unprocessed condition.

For the MP binary mask estimation, we first computed the impulse responses of each gammatone filter and normalized them to have a unit norm. With a dictionary made up of all time-shifts of the impulse responses, we ran MP on the full mixture signal until all remaining coefficients fell below a frequency-dependent threshold. The frequency dependency was implemented to encourage coefficients in bands where speech energy is low. We heuristically chose the coefficient thresholds to scale by a ratio linearly spaced between one-fourth and one. Therefore, the "threshold" specified from this point forward is the coefficient threshold at the lowest frequency band, the coefficient threshold at the highest frequency band is one-fourth of the "threshold," and the coefficient thresholds at the bands in between are linearly spaced between the "threshold" and one-fourth of the "threshold".

Conceptually, we converted the final set of MP coefficients into a binary mask as follows (actual implementation was more efficient). For each of the chosen coefficients, we computed the filter response magnitudes to the corresponding time-shifted gammatone impulse response. Then we assigned the mask a value of one if the response magnitude in the corresponding T-F unit was greater than 1% of the maximum response.

To prevent a rapidly fluctuating mask (particularly in the higher-frequency bands where gammatones are very short in duration), we

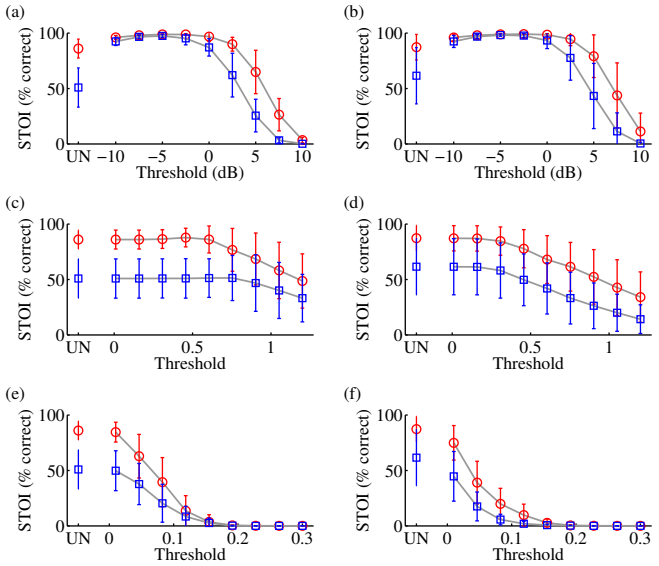

Fig. 2. Average predicted intelligibility in percent words correct (using STOI) when the masks are applied to speech corrupted by pink noise [left column] and realistic noise (babble, car, street, and train) [right column] at 0 dB [red circles] and -5 dB [blue squares] input SNRs for a range of thresholds: (a-b) IBM, (c-d) the MP mask, and (e-f) the FT mask. UN indicates the unprocessed condition. Error bars indicate plus and minus one standard deviation.

post-processed each band of the mask 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, we set the mask to one at all time samples in the frame.

For the FT binary mask estimation, we computed filter response magnitudes for the mixture signal, and then assigned the mask a value of one if the energy in the corresponding T-F unit was greater than or equal to the threshold.

5. RESULTS

We used the short-time objective intelligibility (STOI) measure, which was designed to maintain high correlation with subjective intelligibility of noisy and T-F-masked noisy speech [21], to predict intelligibility outcomes for the mask estimators (Fig. 2). We converted STOI values to predictions of the percentage of words correct using the nonlinear mapping for the IEEE Corpus (Table II in [21]). Predictions for the IBM are consistent with results in the literature. However, STOI does not predict the MP mask to increase intelligibility to the same degree. For the case of pink noise, STOI does predict slight increases 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 (no significance testing). With realistic noise, MP 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 to degrade when the FT binary mask is applied in the presence of realistic noise. In the case of pink noise, STOI predicts the FT mask to maintain intelligibility at very low thresholds but never improve it.

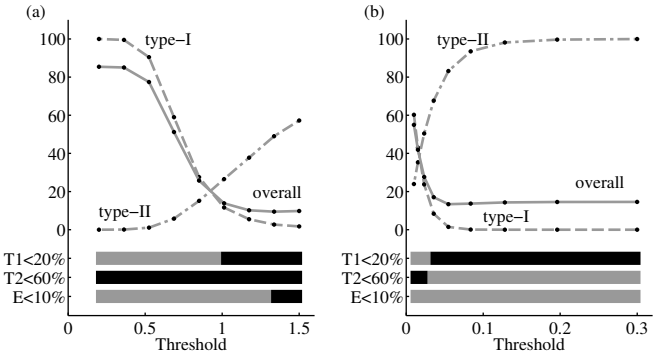

Fig. 3. Binary mask error rate across 16 sentences that have been corrupted with pink noise at -5 dB for (a) the MP mask and (b) the FT mask for a range of thresholds. The thick bars near the bottom indicate when each of the criterion are [black] satisfied and [grey] not satisfied.

6. DISCUSSION

Given that few practical algorithms exist for estimating the IBM in real-world applications, the MP mask is a step in the right direction. However, STOI does not predict that the current MP mask algorithm will increase intelligibility to the same degree as the IBM. We investigated the MP mask further by looking more closely at the type of errors that it makes.

Mask estimators can make one of two types of errors. Type-I errors occur when masker-dominated T-F units are incorrectly labeled target-dominated (i.e., false alarm or false positive), and type-II errors occur when target-dominated T-F units are incorrectly labeled masker-dominated (i.e., miss or false negative). With these definitions, we treat the ground truth as the IBM with a 0 dB threshold. Based on the work presented by Li and Loizou [22], we know that when only uniformly random type-I errors are present, type-II error rates as high as 60% still yield high intelligibility, and when only uniformly random type-II errors are present, type-I error rates as low as 20% are detrimental to intelligibility. Li and Loizou also demonstrated that intelligibility was greater than 90% when overall error rates for uniformly random error (both type-I and type-II together) were at or below 10%. We show in Fig. 3 average error rates for a range of MP and FT masks using the 16 speech sentences corrupted with pink noise at -5 dB SNR. If we were to take into account the conditions set forth by Li and Loizou and assume that meeting the criteria jointly leads to high intelligibility, we would expect a mask that satisfies all three criteria to yield high intelligibility. The FT mask algorithm fails to meet all three criteria jointly. Therefore, we expect to see the poor intelligibility that STOI predicts. However, the MP mask algorithm with a threshold at or around 1.3 meets all three criteria jointly, and therefore, we expect STOI to predict high intelligibility.

Since STOI predictions conflict with this notion, we look more closely at the errors themselves rather than just the *rate* of error for an example sentence in Fig. 4. We compare three masks against the IBM with a 0 dB threshold: (a) the IBM with a -10 dB threshold, (b) the MP mask with a 1.3 threshold, and (c) a simulated mask with uniformly distributed random error. Visually, it is easy to see that the first mask contains a high number of false positives relative to the reference. With regard to the second and third masks, false positives and false negatives are grouped together in the former but randomly distributed in the latter.

Again assuming that the IBM with a 0 dB threshold is ground truth, we can compute error rates for these individual masks: IBM

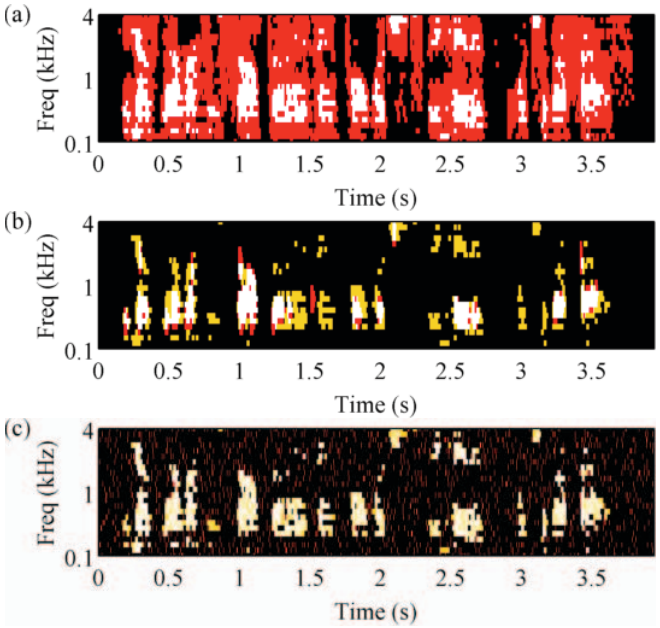

Fig. 4. Time-frequency representations of false positives (red), false negatives (yellow), correct positives (white), and correct negatives (black) relative to the IBM with a 0 dB threshold for: (a) the IBM with a -10 dB threshold, (b) the MP mask, and (c) a simulated mask with uniformly distributed random error.

with a -10 dB threshold (55% type-I, 0% type-I, 47% overall, STOI predict 91% correct), MP (2% type-I, 59% type-II, 9% overall, STOI predict 8% correct), and a simulated uniformly distributed random error case (5% type-I, 26% type-II, 8% overall, STOI predicts 74% correct). First, we point out that 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, we designed the simulated mask to have low type-I error, an overall error rate less than 10%, and type-II error less than 60% (the same criterion we used to design the MP mask). Even though the error rates between the second and third masks are similar, 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 locally. Without listener studies, we cannot be sure of the effect that the distribution of the errors has on the resulting intelligibility.

It is our hypothesis that structured errors influence intelligibility differently than randomly distributed error. As an additional example, we consider another case: IBM with a 2.5 dB threshold (0% type-I, 51% type-II, 7% overall, STOI predicted 41% correct). According to Li and Loizou's study with randomly distributed error [22], we would expect high intelligibility in this case given that all three criterion on the error rates are met. Instead, STOI predicts low intelligibility. It seems that acceptable type-I and type-II error levels are actually reversed for structured error that is grouped in T-F regions as compared to uniformly random error. To summarize, type-II errors are more tolerable when the error in question is randomly distributed, but type-I errors are more tolerable when the error in question is structured in T-F groups. It remains to be seen with listener studies what the allowable overall error rate is for non-randomly distributed error types with interacting type-I and type-I errors.

Performance is very sensitive to the parameters which control the conversion from MP coefficients to the binary mask. Future work will include a more thorough investigation of the optimal method to do this conversion. Furthermore, we will likely need to take into account higher-order statistical structure in order to more accurately distinguish the target speech from challenging interfering signals. To accomplish this, we may need to use sparse approximation alternatives to MP or models beyond sparsity.

7. CONCLUSIONS

We have demonstrated that sparse approximation is a promising direction for binary mask estimation because it enforces a meaningful structure on the binary T-F mask that makes it possible to decrease estimation errors and maintain intelligibility. However, we cannot resolve specific design criteria without listener studies. Although the algorithm presented here is a promising proof of concept, it employs a non-causal estimator. A key factor going forward is altering this approach to use more realistic frame-based (causal) computations.

8. REFERENCES

[1] D Brungart, P Chang, B Simpson, and D Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[2] D Wang, U Kjems, M Pedersen, J Boldt, and T Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[3] M Anzalone, L Calandruccio, K Doherty, and L Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[4] D Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[5] U Kjems, J Boldt, M Pedersen, T Lunner, and D Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[6] G Kim, Y Lu, Y Hu, and P Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[7] S Mauger, P Dawson, and A Hersbach, "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 327–336, 2012.

[8] K Wójcicki and P Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2904–2913, 2012.

[9] F Chen and P Loizou, "Impact of SNR and gain-function over- and under-estimation on speech intelligibility," *Speech Communication*, vol. 54, no. 2, pp. 272–281, 2012.

[10] I Brons, R Houben, and W Dreschler, "Perceptual effects of noise reduction by time-frequency masking of noisy speech," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2690–2699, 2012.

[11] P Jančovič, X Zou, and M Köküer, "Speech enhancement based on Sparse Code Shrinkage employing multiple speech models," *Speech Communication*, vol. 54, pp. 108–118, 2012.

[12] C Sigg, T Dikk, and J Buhmann, "Speech Enhancement Using Generative Dictionary Learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, 2012.

[13] J Sang, G Li, H Hu, M Lutman, and S Bleeck, "Supervised Sparse Coding Strategy in Cochlear Implants," in *InterSpeech*, Florence, Italy, 2011, pp. 1–4.

[14] J Sang, H Hu, G Li, M Lutman, and S Bleeck, "Supervised sparse coding strategy in hearing aids," in *2011 IEEE 13th International Conference on Communication Technology (ICCT)*, 2011, pp. 827–832.

[15] Y Karklin, C Ekanadham, and E Simoncelli, "Hierarchical spike coding of sound," in *2012 Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, 2012.

[16] E Smith and M Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[17] S Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2 edition, 1999.

[18] E Smith and M Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation*, vol. 17, no. 1, pp. 19–45, 2005.

[19] J Garofolo, L Lamel, W Fisher, J Fiscus, D Pallett, N Dahlgren, and V Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[20] V Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica united with Acustica*, 2002.

[21] C Taal, R Hendriks, R Heusdens, and J Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, 2011.

[22] N Li and P Loizou, "Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.