Aalborg Universitet

AALBORG UNIVERSITY

Error-Correction of Binary Masks using Hidden Markov Models

Boldt, Jesper; Pedersen, Michael Syskind; Kjems, Ulrik; Christensen, Mads Græsbøll; Jensen, Søren Holdt

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2010.5495182

Publication date:
2010

Document Version
Accepteret manuscript, peer-review version

Link to publication from Aalborg University

Citation for published version (APA):
Boldt, J., Pedersen, M. S., Kjems, U., Christensen, M. G., & Jensen, S. H. (2010). Error-Correction of Binary Masks using Hidden Markov Models. *I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings*, 4722-4725. https://doi.org/10.1109/ICASSP.2010.5495182

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: December 05, 2025

ERROR-CORRECTION OF BINARY MASKS USING HIDDEN MARKOV MODELS

Jesper Bünsow Boldt[1,2]*, Michael Syskind Pedersen*[1]*, Ulrik Kjems*[1]*,*
Mads Græsbøll Christensen[2]*, Søren Holdt Jensen*[2]

[1]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[2]Department of Electronic Systems, Aalborg University, DK-9220 Aalborg Øst, Denmark
{jeb,msp,uk}@oticon.dk,{mgc,shj}@es.aau.dk

ABSTRACT

Binary masking is a simple yet efficient method for source separation, and certain binary patterns have shown to give a large improvement in speech intelligibility. These binary patterns are difficult to calculate in real-life applications because they require the clean target sound to be available. Accepting that the calculated binary mask will contain errors when calculated using noisy speech, we propose a method for error-correction based on a hidden Markov model of the error-free *target binary mask*. If the method is used in applications like hearing aids or cochlear implants, the complexity must be kept low. This requirement limits the achievable performance, but the results show that it is possible to correct errors in the binary mask and reduce noise energy at the expence of a loss of target energy.

***Index Terms*—** Binary masks, target binary mask, hidden Markov model, speech intelligibility, error-correction.

1. INTRODUCTION

For source separation, time-frequency masking has been widely used [1]. Basically, time-frequency masking is to apply a time-varying and frequency-dependent gain to a signal in a number of frequency channels. However, the main interest in time-frequency masking is not the application of the gain but how to calculate the gain at different times and frequencies. This gain pattern is referred to as the *mask*, and if the mask only contains zeros and ones, the method is referred to as *binary masking*. The interest in binary masking in applications like hearing aids and cochlear implants can be explained by its simplicity and efficiency. Simplicity, because the decision is to either keep the time-frequency unit or remove it, and efficiency with regards to speech intelligibility as shown in the several studies. In [2, 3] the *ideal binary mask* was applied to noisy speech showing a high increase in intelligibility measured by subjective listening tests. This ideal binary mask was further studied in [4] along with the *target binary mask* - both showing a large increase in intelligibility when applied to noisy speech. The target binary mask TBM is calculated by comparing the time-frequency representation of the target sound $\mathbf{T}$ with the long term average energy of the target speech $\mathbf{r}$

$$\mathrm{TBM}_{k,\tau} = \begin{cases} 1, \text{ if } \frac{\mathbf{T}_{k,\tau}}{\mathbf{r}_k} > \mathrm{LC} \\ 0, \text{ otherwise} \end{cases}, \qquad (1)$$

where LC is the local SNR criterion, τ the time index, k the frequency channel, and subscripts refer to the elements in the matrices TBM and $\mathbf{T}$ or elements in the vector $\mathbf{r}$. The LC value controls the density of the TBM as measured by the percentage of ones in the binary mask within speech intervals, and high intelligibility was achieved with densities between 20% and 60% in [4].

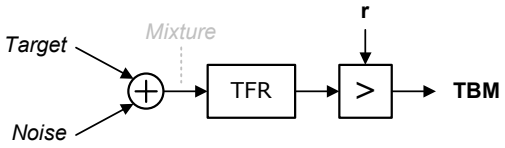

Fig. 1. Setup for calculation of the target binary mask (TBM). The block named TFR calculates the time-frequency representation of the mixture which are compared to the long term average speech energy $\mathbf{r}$ to get the TBM.

The gain in intelligibility makes the TBM interesting in applications where loss of intelligibility is a problem, e.g., for hearing impaired in noisy conditions. The classic approaches to this problem have been examined in [5], where different speech enhancement algorithms are evaluated using normal-hearing listeners, but only a single algorithm in a single noise condition improved intelligiblity significantly. More promising methods exists in the area of source separation (e.g. [6]), but the complexity of these algorithms makes them unusable in low-complexity applications.

The obvious drawback of using the TBM is the requirement of the clean target sound. This requirement is difficult to fulfill in real-life applications. However, estimates of the TBM could be obtained using array processing or noise reduction techniques, which could be followed by an error-correction of the estimated TBM. This two step approach–estimation and error-correction–is not optimal, but it might call for simpler solutions, instead of the optimal solution of estimating the TBM correctly. In this study, we will focus on the error-correction step, well aware that the estimation step is not a trivial problem.

The error-correction will be based on a model of the TBM from a training set of clean speech. This approach assumes that the TBM from different speakers share some common characteristics which can be captured and generalized by the model. The validity of this assumption will be verified in the present work by examining the models ability to identify errors in a TBM calculated from noisy speech. If all errors should be identified, the complexity of the model would make it unusable in low-complexity systems, and a second objective is to examine the efficiency of the error-correction using relative simple models of the TBM.

The setup shown in Figure 1 is used to examine the efficiency of the proposed error-correction method. In this setup, the TBM will be error-free, if no noise sound is present and $\mathbf{r}$ is known. If a noise sound is present, two types of errors will be introduced. *False ones* will be found, if the noise sound increases the energy in time-frequency units to be larger than $\mathbf{r}$, and *false zeros* will be found

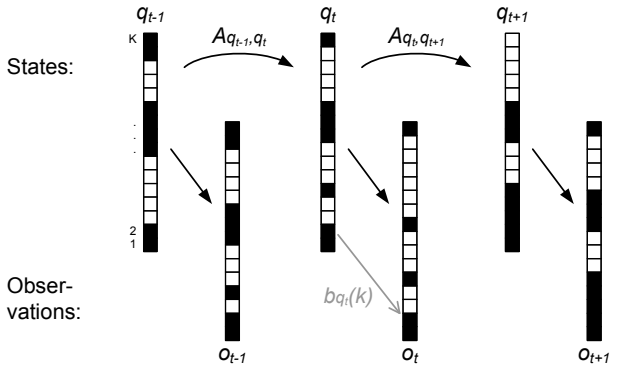

Fig. 2. The structure of the used HMM. At time t, the state q_t generates an observation o_t and change state with probability $\mathbf{A}_{q_t,q_{t+1}}$. The probability of being in state q_t and seeing a one in frequency channel k is defined by $b_{q_t}(k)$ as shown with gray.

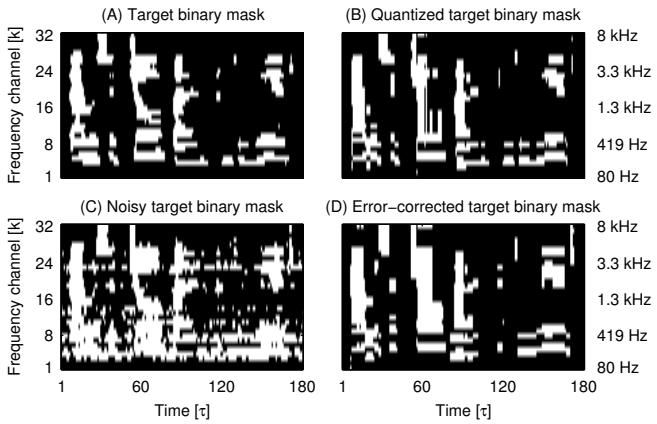

Fig. 3. (A) is the target binary mask from 1.8 s of speech. (B) is (A) quantized using 256 states. (C) is the speech from (A) mixed with speech shaped noise at 0 dB SNR. (D) is (C) after error-correction using a 256 state HMM.

if the target and noise sound cancel each other in certain frequency channels. At high signal-to-noise ratios (SNR), no errors will be introduced in the TBM, but as the SNR decreases the number of false ones will increase. Ultimately, the binary mask will become an all-one mask, but in this situation error-correction is not useful, although forcing some frequency channels to zero could decrease the total number of false units. The expectation is that at intermediate SNRs, when the number of errors are comparable to the number of correct time-frequency units, it is possible to reduce the total number of errors.

2. BINARY MASK MODEL

The error-correction is based on a hidden Markov model (HMM) [7] of the TBM. The HMM is a widely used statistical model for pattern recognition and speech processing and is particularly suited to model time-series with time-varying statistical properties. In the HMM, the hidden layer contains a number of states, which, at each time increament, can change and generate an observation, and from a sequence of observations, the most probable sequence of states can be calculated using the Viterbi algorithm [7].

In the present application, the observations will be the noisy TBM, the states will be the noise-free TBM, and the error-correction will be the step of calculating the most probable noise-free TBM from the noisy TBM using the Viterbi algorithm. The states in the used HMM are binary column vectors with size $K \times 1$, where K is the number of frequency channels as seen in Figure 2. Each state in the HMM represents the TBM at a single time index τ, and the probability of changing state is described by the state-transition probability matrix $\mathbf{A}$, where the element $\mathbf{A}_{i,j}$ is the probability of changing from state i to j. If the TBM is described by a limited number of states N, errors will be introduced as seen in Figure 3. We refer to this process as quantization of the binary mask. In each state j, the observation probability $b_j(k)$ determines the probability of a one in each of the K frequency channels. If we assume that the target sound and noise sound are independent and do not overlap in time, the observation probability in state j is given by

$$b_j(k) = b_j^T(k) + b^N(k) - b_j^T(k)b^N(k), \tag{2}$$

where $b_j^T(k)$ is the probability of a one generated by the target sound, and $b^N(k)$ is the probability of a one generated by the noise sound. If the N states could describe the TBM without quantization errors, the element $b_j(k)$ would identical to the element at frequency k in the binary column vector in state j, but because of the quantization $b_j^T(k)$ will not be binary. If d columns from the training data are quantized to the same state j, and c out of the d columns have a one at frequency index k, we see that $b_j^T(k) = c/d$. The probability of a one generated by the noise $b^N(k)$ is independent of the states but dependent on the SNR, so $b^N(k)$ will be estimated from a short segment of the noise sound prior to the error-correction. The last parameter in the HMM is the initial state distribution which is chosen to be the state with an all-zero column vector.

3. TRAINING

To find the parameters for the HMM, speech from the EUROM corpus was used [8]. First, the TBM was calculated from 36 minutes of speech created by 4 male and 4 female speakers normalized to have equal energy. To calculate $\mathbf{T}$ in (1), a 32 band Gammatone filterbank [9] with centerfrequencies from 80 Hz to 8000 Hz equally spaced on the ERB (equivalent rectangular bandwidth) scale was used. The energy from each filterbank channel was divided into 20 ms frames with 10 ms overlap, and an LC value of 0 dB was used in (1). The long term average energy $\mathbf{r}$ used in (1) was the long term average energy of the 8 speakers and not the long term average energy of the individual speakers as used in [4]. The 36 minutes of speech produced a training TBM with size 216000×32 from which N states were found while minimizing the quantization error measured by the total number of false ones and false zeros. This quantization was done using the K-mode algorithm which is similar to the well-known K-means algorithm but suitable for clustering binary data [10]. From the quantized TBM, the state-transition probabilities $\mathbf{A}_{i,j}$ was calculated by counting the number of state changes from i to j and divide by the total number of visits in state i. To find $b_j^T(k)$, the columns in the training data that were quantized to the same state j were collected and the probability of a one in each of the K frequency channels was calculated. $b^N(k)$ was calculated online using the first 5 seconds of the noisy speech where no target speech is present. From the binary mask obtained with the noise, $b^N(k)$ can be calculated by counting the number of ones in each frequency channel k and divide by the length of the binary mask.

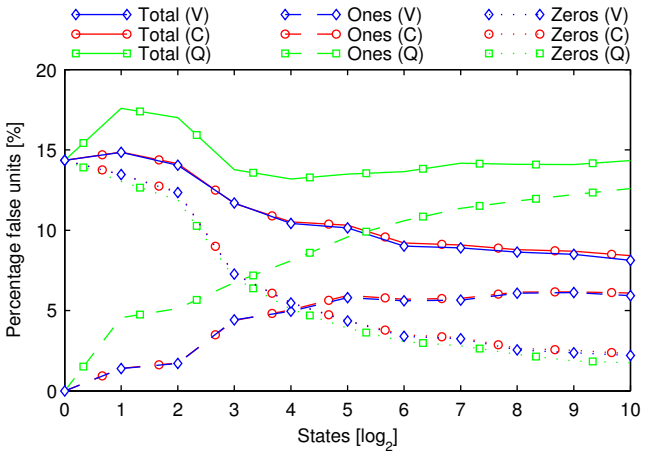

Fig. 4. False units in the error-corrected target binary mask as a function of number of states in the HMM. (V) is error-correction using the Viterbi algorithm, (C) is error-correction using a causal Viterbi algorithm, and (Q) is simple quantization of the noisy TBM. The percentages of false units before error-correction were 0.2% false zeros, 15.9% false ones, and 16.1% in total.

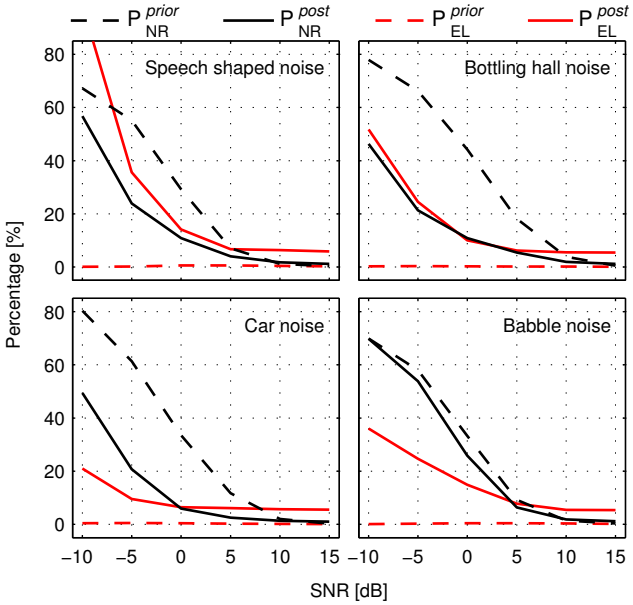

Fig. 5. The noise residue P_{NR} and the loss of target energy P_{EL} are shown before and after the error-correction. The HMM used for error-correction has 256 states.

4. EVALUATION

To evaluate the proposed method for error-correction of the TBM, two simulations were carried out. The first simulation shows the relation between the number of states and the performance of the error-correction measured by the number of correct time-frequency units. The second simulation shows the performance at different SNRs and noise sounds measured by the loss of target energy and the remaining noise in the output prior or post the error-correction. In both simulations 10 sentences from a male and female speaker were used, and the two speakers were not part of the training data.

In the first simulation, the sentences were mixed with speech shaped noise at 0 dB SNR. An HMM with a varying number of states was trained as described in Section 3 and used to correct errors in the noisy TBM. In Figure 4, the percentage of false time-frequency units is shown after the error-correction, and for comparison the number of false units are also shown for the quantized noisy TBM and error-correction using a causal Viterbi algorithm. All percentages are calculated relative to the total number of time-frequency units in the binary mask. The percentages of false units in the TBM before error-corrections were 0.2% false zeros, 15.9% false ones, and 16.1% in total. As the number of states increases in Figure 4, the number of false ones increases, whereas the number of false zeros decreases. At the lower limit with a single state in the HMM, this single state will be the all-zero column vector, making it impossible to have false ones in the error-corrected binary mask. When the number of states is increased, the states have low densities and limit the maximum number of false ones, but as the number of states is further increased the number of false units levels off. Using 1024 states, the number of false units in total is 8.1% - a reduction of 8 percentage point compared to the noisy TBM, but the reduction of false ones is achieved at the cost of an increase in false zeros relative to the noisy TBM. The Viterbi algorithm uses previous, current and future columns from the noisy TBM to calculate the best state sequence. In low-delay applications, the dependency on the future is critical and for comparision a causal Viterbi has been implemented. This algorithm finds the best state sequence using only the previous and current columns in the noisy TBM, and, as seen in Figure 4, this modification does not reduce performance significantly.

In the second simulation, the performance at -10 dB to 15 dB SNR was examined using a 256 state HMM trained as described in Section 3. The sentences were mixed with four different noise types: speech shaped noise, a high-frequency sound from a bottling hall, a low-frequency sound from the interior of a car, and a 7 speaker babble noise. Performance was measured using the *percentage of energy loss* and the *percentage of noise residue* [11]:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \quad (3) \qquad P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}. \quad (4)$$

$I(n)$ is the resynthesized sound using the TBM, $O(n)$ is the resynthesized output prior or post error-correction, $e_1(n)$ is the sound found in $I(n)$ but not in $O(n)$, and $e_2(n)$ is the sound found in $O(n)$ but not in $I(n)$. From the results in Figure 5, we see a similar behavior with the four different noise types. At low SNR the percentage of unwanted noise energy P_{NR}^{prior} is very high for the noisy TBM before error-correction. As the SNR is increased, the number of false ones in the TBM is decreased resulting in a lower percentage of unwanted noise energy. Ultimately, when SNR is further increased, P_{NR}^{prior} reduces to 0% because no false ones is found in the noisy TBM. The noise residue after error-correction P_{NR}^{post} shows that the noise energy is reduced at SNRs below 10 dB but slighly increased at SNRs around 15 dB. This increase shows that error-correction of an error-free TBM can introduce false ones because of the limited number of states in the HMM. The measure P_{EL}^{prior} shows that loss of target energy using the noisy TBM is close to 0%, because very few false zeros are found in the noisy TBM before error-correction. When error-correction is introduced the loss of target energy is increased as shown by P_{EL}^{post}: At low SNRs, we find a significant loss of target energy, but as the SNR increases this loss is reduced and levels off at around 8%. The lower

limit of P_{EL}^{post} at 8% has the same explanation as the increase in P_{NR}^{post} at high SNRs: The limited number of states in the HMM will increase the number of false ones and zeros in the TBM, when the noisy TBM contains very few errors. For all four noise types expect the babble noise, the best performance is found around $0-5$ dB SNR, when the error-correction reduces the noise energy more than the target energy. Listening to the processed mixtures of speech and noise before and after error-correction confirms this finding.

5. DISCUSSION

The results confirm that a model of the TBM can be build and used to correct errors in the noisy TBM. As seen in the two simulations, the reduction of false ones comes with a price of an increase in false zeros. Even though the relation between the percentage of false time-frequency units and intelligibility has been established in [12], it is difficult to use their results to determine the intelligibility of the noisy or error-corrected TBM. The errors in [12] are uniformly spread in time and frequency and that assumption is not correct for the errors in the present study. In [12], they find that false ones reduce intelligibility more than false zeros, but one must assume that the location of errors and the noise type have a significant impact on intelligibility, e.g. if the false zeros are found at onsets in the target speech.

Knowing the relation between false time-frequency units and intelligibility could change the model parameters. If it was known that false ones would reduce intelligibility more than false zeros, the model could be changed to allow more false zeros than false ones. This weighting would make it possible to adjust the relative level of P_{NR}^{post} and P_{EL}^{post}. Furthermore, the errors impact on intelligibility is probably frequency dependent making it useful to reduce quantization error at certain frequencies at the expense of increases at other frequencies.

The increase in false zeros from the error-correction gave a loss of target energy. This drawback could be reduced, if a lower LC value was used in (1) to calcualte the TBM. The results in [4] show that high intelligibility can be obtained for a range of LC values, so the loss of target energy does not necessarily cause a loss in intelligibility. However, lowering the LC value will also increase the number of false ones in the binary mask.

An interesting question to consider is if the performance of the error-correction will continue to improve with an increasing number of states. Using more states will reduce the quantization error, but errors will be inevitable. False ones in the TBM can make a wrong sequence of states more probable than the correct sequence and similar with false zeros. This limitation is a drawback of working in the binary domain, because the amount of information about the target and noise sound is greatly reduced compared to the time-frequency representation of the sounds from which the TBM is calculated.

Another limitation of the model is the use of multiple speakers and the long term average energy $\mathbf{r}$ in (1) not being adjusted to the individual speaker as in [4]. Multiple speakers give a model of the TBM that is more general but less precise for the individual speakers. This problem might be solved by adjusting $\mathbf{r}$ to the speaker, and in that way obtain a TBM that is less different between speakers. If a speaker independent TBM could be obtained, the model would be stronger but this is not easily obtained.

More complex models, e.g., factorial HMMs, could also make the error-correction more efficient. However, the complexity of the model should be considered carefully with respect to the complexity of the signal or pattern. The binary mask is a simplified pattern compared to the time-frequency representation, and a very complex model of this simplified pattern might not be optimal. Instead, models of the time-frequency representation should be considered. This statement is also relevant for the present work, because, even though the complexity of the model is low, the computational cost of using the large number of states can be a problem in applications like hearing aids.

6. CONCLUSION

We have proposed a method for error-correction of the TBM calculated from noisy speech. The method is based on a HMM and trained on noise-free examples of the TBM, and it has been shown that model can correct errors, although the reduction of false ones has the drawback of increasing the number of false zeros. Knowing that this error-correction can be succesfull makes algorithms for estimation of the TBM more interesting and useful in real-life applications like hearing aids and cochlear implants. The model used in this study can be further improved, e.g. by weighting of different time-frequency units, but the model can also be useful for similar problems involving binary patterns obtained from noisy observations.

7. REFERENCES

[1] 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.

[2] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[3] N. Li and P. C. Loizou, "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. EL59–EL64, 2008.

[4] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[5] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 1 edition, 2007.

[6] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," *Proc. Interspeech*, vol. 1, pp. 97–100, 2006.

[7] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[8] D. Chan et al., "EUROM - a spoken language resource for the EU," in *Proc. Eurospeech*, 1995, vol. 1, pp. 867–870.

[9] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS final report, part b: Implementing a gammatone filterbank," *Rep. 2341, MRC Applied Psychology Unit.*, 1988.

[10] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[11] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[12] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, 2008.