ESTIMATION OF THE IDEAL BINARY MASK USING DIRECTIONAL SYSTEMS

Jesper Bünsow Boldt[1,2], Ulrik Kjems[2], Michael Syskind Pedersen[2], Thomas Lunner[3], DeLiang Wang[4]

[1]Department of Electronic Systems, Aalborg University, DK-9220 Aalborg Øst, Denmark
[2]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[3]Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark
[4]Department of Computer Science and Engineering & Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277, USA
email: {jeb, uk, msp, tlu}@oticon.dk, dwang@cse.ohio-state.edu

ABSTRACT

The ideal binary mask is often seen as a goal for time-frequency masking algorithms trying to increase speech intelligibility, but the required availability of the unmixed signals makes it difficult to calculate the ideal binary mask in any real-life applications. In this paper we derive the theory and the requirements to enable calculations of the ideal binary mask using a directional system without the availability of the unmixed signals. The proposed method has a low complexity and is verified using computer simulation in both ideal and non-ideal setups showing promising results.

Index Terms— Time-Frequency Masking, Directional systems, Ideal Binary Mask, Speech Intelligibility, Sound separation

1. INTRODUCTION

Time-frequency masking is a widely used technique for speech and signal processing used in automatic speech recognition [1], computational auditory scene analysis [2], noise reduction [3, 4], and source separation [5, 6, 7, 8]. The technique is based on time-frequency (T-F) representation of signals and makes it possible to utilize the temporal and spectral properties of speech and the assumption of sparseness of speech. An important quality of T-F masking is the availability of a reference mask, which defines the maximum obtainable speech intelligibility for a given mixture. This *ideal binary mask* (IBM) [9] has recently been demonstrated to have large potential for improving speech intelligibility in difficult listening conditions [10, 4, 3]. To calculate the IBM, the unmixed signals must be available, which is a a requirement rarely met in any real-life application. However, the significant increase in speech intelligibility by the IBM makes it a valuable goal for T-F algorithms trying to increase speech intelligibility. The T-F representation is obtained using e.g. the short-time Fourier transform or a Gammatone filterbank [11], and the IBM is calculated by comparing the power of the target signal to the power of the masker (interfering) signal for each unit in the T-F representations:

$$\mathrm{IBM}(\tau,k) = \begin{cases} 1, \text{ if } \dfrac{\mathbf{T}(\tau,k)}{\mathbf{M}(\tau,k)} > \mathrm{LC} \\ 0, \text{ otherwise} \end{cases}, \tag{1}$$

where $\mathbf{T}(\tau,k)$ is the power of the target signal, $\mathbf{M}(\tau,k)$ is the power of the masker signal, LC is a local SNR criterion, τ the time index, and k the frequency index. The LC value is the threshold for classifying the T-F unit as target or masker and determines the amount of target and masker signal in the processed signal, if the binary mask is applied to the mixture. In computational auditory scene analysis (CASA), an LC value of 0 dB is commonly used, but recent studies have shown that a certain range of LC values different from zero provides the same major improvement in speech intelligibility [10, 3].

In this paper we show that it is indeed possible to calculate the IBM without the availability of the unmixed signals. This is made possible with the proposed method and the required theory and constraints are derived. The proposed method has a very low complexity and is based on a first-order differential array. To verify the method and document the theory, computer simulations are performed: First, in the ideal situation where all constraints are met, and subsequently in situations where one or more constraints are not met. These simulations verify the precision of the method in the ideal situations, and the robustness of the method in non-ideal situations.

2. IBM ESTIMATION

The proposed method is based on two first-order differential arrays (cardioids) pointing in opposite directions. One target source and one masker source are present and separated in space as shown in Figure 1. We assume that the directional patterns and the azimuths of the two sources are known. If the spacing between the two microphones in the first-order differential array is much smaller than the acoustic wavelength, the output can be approximated by [12]:

$$C_T(f) \approx G(f)\left(a_0 T(f) + a_1 M(f)\right) \tag{2}$$
$$C_M(f) \approx G(f)\left(b_0 T(f) + b_1 M(f)\right), \tag{3}$$

where f is the frequency, $G(f)$ is a high-pass system, $T(f)$ is the target signal, $M(f)$ is the masker signal, and a_0, a_1, b_0, b_1 are directional gains for the target and masker signal as shown in Figure 1. To obtain the T-F representations of $C_T(f)$ and $C_M(f)$ the two signals are further processed as shown in Figure 2: Filtering through a K-point filterbank, squaring the absolute value, low-pass filtering, and downsampling by a factor P. Assuming that $T(f)$ and $M(f)$ are uncorrelated, the four steps result in the two directional power signals:

$$\mathbf{D}_T(\tau,k) = |G(k)|^2\left(a_0^2\mathbf{T}(\tau,k) + a_1^2\mathbf{M}(\tau,k)\right) \tag{4}$$
$$\mathbf{D}_M(\tau,k) = |G(k)|^2\left(b_0^2\mathbf{T}(\tau,k) + b_1^2\mathbf{M}(\tau,k)\right), \tag{5}$$

where $\mathbf{T}(\tau,k)$ and $\mathbf{M}(\tau,k)$ are the powers of the target and masker signals, respectively. To estimate the IBM using the two directional

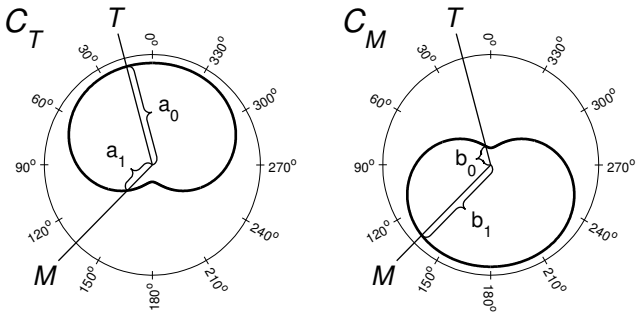

Fig. 1. The directional patterns of the two first-order differential arrays. C_T points towards the target signal T, and C_M points towards the masker signal M. The directional gains a_0, a_1, b_0, and b_1 are functions of the azimuths of the two sources T and M.

power signals (4, 5), we change (1) to

$$\widehat{\mathrm{IBM}}(\tau,k) = \begin{cases} 1, \text{ if } \dfrac{\mathbf{D}_T(\tau,k)}{\mathbf{D}_M(\tau,k)} > \mathrm{LC}' \\ 0, \text{ otherwise} \end{cases}, \tag{6}$$

where LC′ is the applied local SNR criterion derived in the next section, and $\widehat{\mathrm{IBM}}$ is the estimate of the IBM.

2.1. The relation between LC and LC′

To estimate the IBM with the directional system using (6), the LC′ value must be derived from the LC value used in the definition of the IBM (1). Leaving out the time and frequency indices in the directional signals from (4, 5) we get, using (6):

$$\frac{a_0^2\mathbf{T} + a_1^2\mathbf{M}}{b_0^2\mathbf{T} + b_1^2\mathbf{M}} > \mathrm{LC}' \;\Leftrightarrow\; \frac{\mathbf{T}}{\mathbf{M}} > \frac{b_1^2\mathrm{LC}' - a_1^2}{a_0^2 - b_0^2\mathrm{LC}'}. \tag{7}$$

To allow this rearrangement, we introduce the constraints

$$a_0^2 - b_0^2\mathrm{LC}' > 0 \quad \text{and} \quad b_1^2\mathrm{LC}' - a_1^2 > 0, \tag{8}$$

which guarantee that $\mathbf{T}/\mathbf{M} > 0$ and prevent the target and masker from being interchanged. A prerequisite for estimating the IBM is that C_T captures more target signal than masker signal, and C_M captures more masker signal than target signal. Otherwise, the binary mask will be inverted. Using the definition of the IBM from (1) in combination with (7) we obtain

$$\mathrm{LC} = \frac{b_1^2\mathrm{LC}' - a_1^2}{a_0^2 - b_0^2\mathrm{LC}'} \Leftrightarrow \tag{9}$$

$$\mathrm{LC}' = \frac{a_0^2\mathrm{LC} + a_1^2}{b_0^2\mathrm{LC} + b_1^2}. \tag{10}$$

Since we can express LC′ in terms of LC, we can actually estimate the IBM without having the unmixed sounds available, if the directional gains are known.

2.2. The asymptotes of LC′

If the directional gains are known, the LC′ value can be calculated from the wanted LC value using (10). If the directional gains are unknown, a fixed LC′ must be used in (6), and the LC value will change depending on the location of the sources (9). Combining the two constraints from (8) we get that

$$\frac{a_1^2}{b_1^2} < \mathrm{LC}' < \frac{a_0^2}{b_0^2}, \tag{11}$$

which are the two asymptotes of LC′ as shown in Figure 3. The asymptotes are determined by the amount of target and masker signal captured by C_T compared to C_M. If no target signal is found in C_M, the high asymptote will be at $+\infty$ dB, and if no masker signal is found in C_T, the low asymptote will be at $-\infty$ dB. In the interval bounded by the two asymptotes we find a region where the relation between LC and LC′ becomes approximately linear. In this region, changes of LC′ produce an equal change of LC. However, changes of LC′ near the asymptotes produce very large changes of LC. We refer to this relation as the *sensitivity* of the method. If the sensitivity is high, errors on $\mathbf{D}_T$, $\mathbf{D}_M$, or the directional gains, can have a significant impact on the LC value. The minimum sensitivity is found in the approximately linear regions which should be as large as possible. The asymptotes makes the LC′ be defined for all LC values, whereas the opposite is not true. If the LC′ value used in (6) is below the low asymptote, the mask becomes an all-one mask. If the LC′ is above the high asymptote the mask becomes an all-zero mask.

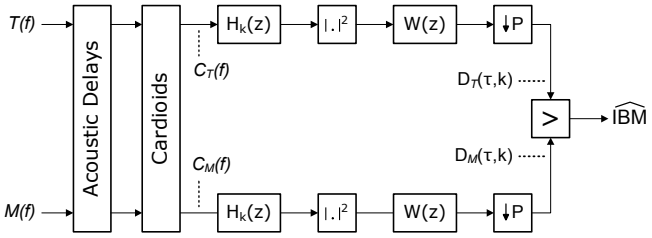

Fig. 2. Blockdiagram for estimation of the ideal binary mask. The acoustic delays model the delay from sources to the microphones in the first-order differential array. $\mathrm{H_k(z)}$ is the k'th analysis filter in the filterbank, $\mathrm{W(z)}$ is a low-pass filter, and $\downarrow\mathrm{P}$ is a decimation. The block labeled $>$ is the implementation of Equation (6).

3. SIMULATIONS

To verify that it is possible to estimate the IBM with the proposed method, a computer simulation was performed showing the precision of the estimate. Furthermore, simulations were done in non-ideal situations to illustrate the robustness of the method. The precision were measured by the number of correct T-F units in the $\widehat{\mathrm{IBM}}$ with respect to the IBM. Two instances of the system shown in Figure 2 were used: The first instance was used to calculate the $\widehat{\mathrm{IBM}}$ and was configured as follows: The acoustic delays were calculated from the azimuth of the two sources using a free-field model [13] with no reverberation. Two microphones were placed with a distance of 1 cm on the line through 0° and 180°, and the distance from the microphones to the sources was 1 m. Two cardioid signals were derived from the microphone signals, and each of the cardioid signals was processed by a 128 band Gammatone filterbank [11] with center frequencies linearly distributed on the ERB frequency scale from 100 Hz to 8000 Hz, each filter having a bandwidth of 1 ERB. The LP filter $W(z)$ was a 20 ms rectangular window followed by a 100 fold decimation corresponding to a 10 ms shift at the used sampling frequency of 20 kHz. The second instance of the system from Figure

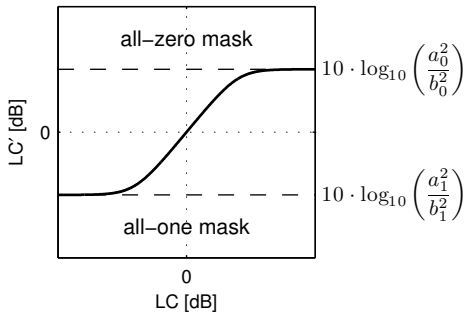

Fig. 3. LC′ as a function of LC. The asymptotes are defined by the directional gains. Using LC′ values outside the region bound by the two asymptotes produce all-one or all-zero masks.

2 was used to calculate the IBM. This instance was equal to the previous without the cardioids. Instead, the target and masker sound were recorded separately by a single microphone located between the microphones used in the previous instance.

In the first simulation, the free-field model was used to calculate the acoustic delays, while the masker source was moved from $180 - 0^\circ$, and the target source was fixed at 30°. The two sources were male and female speech with 0 dB SNR and a duration of 11 seconds. A fixed LC′ value of 0 dB was compared to an adaptive LC′ value calculated using (10) and an LC value of 0 dB.

3.1. Simulation 1

The results from the first simulation are shown in Figure 4. The solid line is the percentage of correct T-F units using an adaptive LC′ value, and the dashed line is LC′ fixed at 0 dB. In both situations we see a high percentage of correct T-F units when the masker azimuth is in the range $180^\circ - 150^\circ$, and the small number of wrong T-F units ($< 2\%$) can be explained by the cardioid filters only used to calculate the $\widehat{\text{IBM}}$.

As the masker source is moved towards the target source, the percentage of correct T-F units decreases faster for the fixed LC′ than the adaptive LC′. At 90° the fixed LC′ has decreased to almost 50% whereas the adaptive LC′ remains above 95%. This decrease is explained by the $\widehat{\text{IBM}}$ becoming an all-one mask which in this case has around 50% correct T-F units. When the masker azimuth is 90° an equal amount of masker signal is captured by C_T and C_M, and the low asymptote in Figure 3 will be at 0 dB. In this situation the 0 dB fixed LC′ value is equal to an LC value of $-\infty$ dB. Moving the masker source further, we see a rapid decrease in correct T-F units for the adaptive LC′, when the masker source passes the target source at 30°. The decrease from above 90% to below 10% correct T-F units is explained by the interchange of target and masker because (11) is not satisfied anymore. If C_T captures more masker than target sound or C_M captures more target than masker sound, the $\widehat{\text{IBM}}$ is the inverse of the IBM with a very low number of correct T-F units.

The small decrease in correct T-F units for the adaptive LC′ value between 180° to 45° can be explained by increased sensitivity of the system. As the masker and target get closer, the two asymptotes from Figure 3 get closer which leads to amplification of the errors introduced by the cardioid filters used for calculating the $\widehat{\text{IBM}}$.

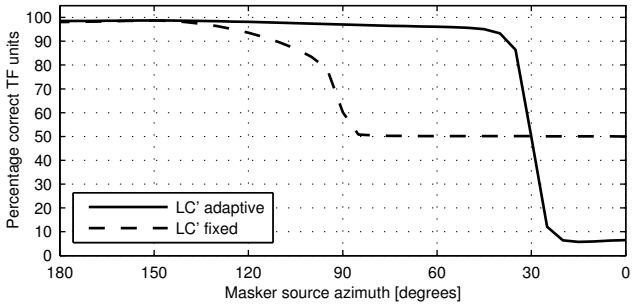

Fig. 4. The percentage of correct T-F units in the $\widehat{\text{IBM}}$ with respect to the IBM. The target was fixed at 30° while the masker was moved from 180° to 0°. The adaptive LC′ value was calculated from the directional gains using an LC value of 0 dB, whereas the fixed LC′ was kept at 0 dB.

3.2. Simulation 2

To further examine the precision and robustness of the proposed method in a non-ideal setup a second simulation was carried out. The setup was identical to simulation 1, except the number of sources and the acoustical delays. One target and three masker sources were present: A male target speaker at 0°, a female masker speaker moving from 180° to 0°, a female masker speaker at 135°, and a male masker speaker at 180°. The speakers were located 2 m from the microphones and the sounds have a duration of 15 seconds. The acoustical delays were the free-field model from simulation 1 and impulse responses from a behind-the-ear (BTE) hearing aid shell on a Head and Torso Simulator (HATS) in three different acoustical environments: Anechoic, low reverberation time (RT_{60}=400 ms), and high reverberation time (RT_{60}=1000 ms). The reverberation time is defined as the time before the room impulse response is decreased by 60 dB.

As in the previous simulation, it is evident from Figure 5 that the percentage of correct T-F units decreases when the moving masker passes 90°. In Figure 4 the fixed LC drops to 50% whereas in Figure 5 the free-field simulation drops to around 72% correct unit. This difference is explained by the two masker sources at 135° and 180° in simulation 2, which prevent the mask from becoming an all-one mask. Compared to simulation 1, where the all-one mask has 50% correct T-F units, the all-one mask in simulation 2 has 34% correct T-F units. Using impulse responses from a hearing aid on a HATS in an anechoic room, the percentage of correct T-F units between 95° and 40° is increased compared to the free-field simulation. This increase is explained by the cardioids being non-ideal and attenuating the moving masker more at these angles. As soon as reverberation is present, the precision of the $\widehat{\text{IBM}}$ decreases. Using impulse responses from the low reverberant room we get around 83% correct units when the moving masker is located at 180°. If the wrong T-F units at this point are divided into wrong ones and wrong zeros with respect to the IBM we find 14% wrong zeros and 19% wrong ones. In other words, the $\widehat{\text{IBM}}$ will remove 14% of the target signal and will retain 19% of the masker signals compared to the IBM if applied to the mixture signal.

4. DISCUSSION

In this paper an important connection between the ideal binary mask and a realizable computation of the binary mask has been estab-

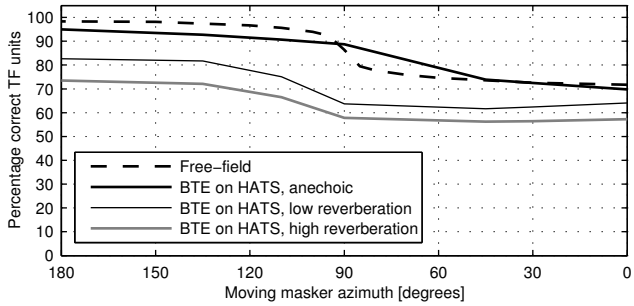

Fig. 5. The percentage of correct T-F units in the $\widehat{\mathrm{IBM}}$ with respect to the IBM. Free-field and impulse responses from a hearing aid shell (BTE) on a HATS in three different acoustical environments were used, and four sources were present: Target at 0°, a moving masker from 180° to 0°, and two fixed maskers at 135° and 180°. The LC$'$ value was 0 dB in all simulations.

lished. To calculate the IBM, the target and masker signals must be available prior to being mixed. This requirement can be relaxed by using a directional system to estimate the IBM, and from (6), we see that the $\widehat{\mathrm{IBM}}$ can be equal to the IBM if only two sources are present, and their directional gains are known. The directional gains are used to calculate the LC$'$ value from the LC value and requires that the directional patterns of the cardioids and the target and masker azimuth are known.

From the first simulation, we find that the proposed method makes it possible to obtain an estimate of the IBM with a very high precission. When the two sources are spatially well separated, the setup with fixed LC$'$ and adaptive LC$'$ both provide a high number of correct T-F units. But as the two sources become closer, the setup with the adaptive LC$'$ shows a significant advantage compared to the fixed LC. The simulation illustrates what happens when the masker source is captured equally by the target and masker cardioid. The binary mask becomes an all-one mask with 50% correct T-F units. The same situation occurs when the target source is captured equally by the two cardioids, and the result is an all-zero mask. The method of varying the LC$'$ value has an advantage over fixating the LC$'$ value, and the target and masker source can become closer before the estimate is degraded significantly.

In the second simulation, we examine the robustness of the proposed method, when conditions are changed from the ideal ones. Introducing more sources and impulse responses from a BTE shell on a HATS in an anechoic room does not undermine the method and a significant increase in speech intelligibility can still be expected from the proposed method. However, a significant decrease in the percentage of correct T-F units is seen when reverberation is introduced, which are agreeable with the results reported using the DUET algorithm in echoic environments [7]. The errors introduced in the estimated binary mask can be divided into two types of errors, and in [3] the wrong ones and wrong zeros are referred to as type I and type II errors, respectively. In their paper, the impact on speech intelligibility of the two types of errors are measured showing that type II errors have a larger impact on speech intelligibility compared to type I errors. This interesting result should be taken into consideration when further developing the proposed method, but the results from [3] can not be used directly to predict speech intelligibility of the method proposed in the present paper. One reason is the difference in setup: We use a Gammatone filterbank whereas a linear filterbank is used in [3]. Another reason is the distribution of errors: It is expected that type II errors scattered uniformly as in [3] will have less impact on speech intelligibility compared to e.g. type II errors placed at onsets in the target sound.

5. CONCLUSION

In this paper we have proposed a method that makes it possible to estimate the ideal binary mask without having the unmixed signals available. If certain constraints are met, the precision of the estimated binary mask is very high, and even if the constraints are not met the proposed method shows promising results having the low complexity of the method in mind. These results establish an important connection between the ideal binary mask and a realizable system for T-F masking, and the precision and robustness of the proposed method in non-ideal conditions makes it very promising for further research and development.

6. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, no. 3, pp. 267–285, 2001.

[2] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis*, Wiley & IEEE Press, Hoboken, New Jersey, 2006.

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

[4] 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.

[5] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *JASA*, vol. 114, no. 4, pp. 2236–2252, 2003.

[6] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. ICA 2004*, Granada, Spain, September 22-24. 2004, pp. 832–839.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[8] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *IEEE Trans. on Neural Networks*, vol. 19, no. 3, 2008.

[9] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Pierre Divenyi, Ed., pp. 181–197. Kluwer, 2005.

[10] 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," *JASA*, vol. 120, no. 6, pp. 4007–4018, 2006.

[11] 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.

[12] G. W. Elko, "Superdirectional Microphone Arrays," in *Acoustic Signal Processing for Telecommunication*, Steven L. Gay and Jacob Benesty, Eds., chapter 10, pp. 181–237. Kluwer Academic Publishers, 2000.

[13] J. Blauert, *Spatial hearing. The Psychophysics of human sound localization*, MIT Press, Cambridge, USA, 1999.