THE IDEAL INTERAURAL PARAMETER MASK: A BOUND ON BINAURAL SEPARATION SYSTEMS

Michael I. Mandel[*]

LabROSA, Dept. Elec. Eng.
Columbia University, NY, NY, 10027
mim@ee.columbia.edu

Daniel P. W. Ellis

LabROSA, Dept. Elec. Eng.
Columbia University, NY, NY, 10027
dpwe@ee.columbia.edu

ABSTRACT

We introduce the Ideal Interaural Parameter Mask as an upper bound on the performance of mask-based source separation algorithms that are based on the differences between signals from two microphones or ears. With two additions to our Model-based EM Source Separation and Localization system, its performance approaches that of the IIPM upper bound to within 0.9 dB. These additions battle the effects of reverberation by absorbing reverberant energy and by forcing the ILD estimate to be larger than it might otherwise be. An oracle reliability measure was also added, in the hope that estimating parameters from more reliable regions of the spectrogram would improve separation, but it was not consistently useful.

Index Terms— Underdetermined source separation, time-frequency masking, ideal binary mask, reverberation

1. INTRODUCTION

Underdetermined mixtures are sound recordings involving more sources than microphones. Humans are adept at solving such problems, having just two ears, but mathematically, they are difficult to separate. One promising separation approach for such mixtures is the generation of time-frequency masks from the interaural parameters of a binaural recording [1, 2]. Reverberation is known to decrease the performance of these algorithms and overcoming this limitation is still an open problem.

The Ideal Binary Mask (IBM), also known as the oracle binary mask, has been proposed as an upper bound on the performance of source separation algorithms that generate binary time-frequency masks [3]. It is "ideal" both in being optimal in terms of signal-to-distortion ratio and in being unattainable from real-world mixtures. This is because it is created using knowledge of the signals before they were mixed. In this paper, we propose a similar upper bound on algorithms that perform time-frequency masking using only point-wise interaural parameters.

This upper bound, which we call the Ideal Interaural Parameter Mask (IIPM), has access to the pre-mixed signals, but creates a time-frequency mask based solely on interaural level and phase differences (ILD and IPD, respectively). All points at a given frequency having a particular ILD and IPD must be either included or excluded from the mask together. By comparing the performance of such an estimator to that of the ideal binary mask, it is possible to determine the separation power of the interaural parameters in reverberation and additional separation performance that must be sought through other means, e.g. monaural source separation, source modeling, dereverberation, etc.

We also make a number of changes to our Model-Based EM Source Separation and Localization (MESSL) system [2] to better compensate for the effects of reverberation and show that the updated model approaches the performance limit of the IIPM. These changes include explicitly modeling reverberation as a separate source, called the "garbage" source, and adding a prior distribution to the ILD parameters for each source based on its estimated location. We also modified MESSL to learn model parameters only from highly reliable regions of the spectrogram, i.e. those regions that have a high direct-to-reverberant ratio (DRR), but found that in general it did not aid separation much.

Several authors have examined the use of supervised learning for source separation, techniques that are related to the IIPM. Perhaps the first, [4] trained classifiers to distinguish between target and interfering sources in anechoic mixtures. Taking an approach similar to the current one, [5] learned a histogram over interaural level and *time* differences for sources in known positions and compared them to a general background model. They used an auditory spectrogram with cross-correlations in each band, whereas we examine the short-time Fourier transform and the phase relationships between individual complex samples.

2. SOURCE SEPARATION SYSTEMS

In this paper, we compare two different source separators, the IIPM and MESSL, which we now describe in detail.

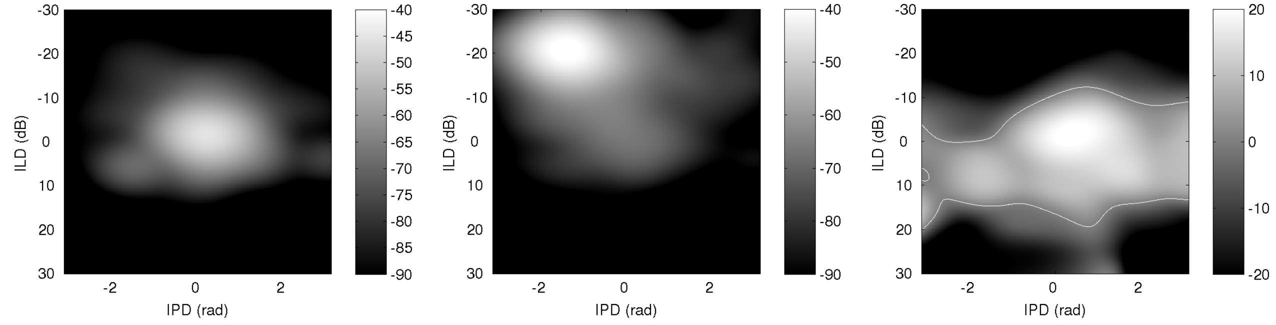
2.1. Ideal Interaural Parameter Mask

The DUET algorithm [1] separates sources by clustering time-frequency points based on their interaural parameters from a recording. It constructs a histogram of interaural parameters, with points weighted by their energy, and then selects each prominent peak in the histogram as the interaural parameters of one source. It then creates a mask for each source that retains only time-frequency points with interaural parameters near the selected peak.

The Ideal Interaural Parameter Mask uses DUET's interaural parametrization to construct a mask similar to the IBM. From knowledge of the separated sources, the IIPM creates models for the interaural parameters weighted by the target and interferer energies. Specifically, it models them with kernel density estimators (KDEs) [6], nonparametric models similar to histograms.

As in DUET, these kernel density estimates are created by weighting the observation at each time-frequency point by the energy of the signal at that point. Let $L(\omega, t)$ and $R(\omega, t)$ be the

[*] This work was funded by NSF grant IIS-0535168 and by EU project AMIDA via ICSI, Berkeley.



(a) (b) (c)

Figure 1: Example kernel density estimates in dB of (a) target energy, (b) interferer energy, (c) energy ratio between target and interferer with lines indicating the decision boundary at 0 dB. From the 4125 Hz band, target at 0° and interferer at 90°.

spectrograms of the observations at the left and right ear. The interaural level difference, measured in dB, $\alpha(\omega,t)$, and the interaural phase difference, $\phi(\omega,t)$ are related to the observations by

$$\tfrac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}. \tag{1}$$

Let $x(\omega,t) = [\alpha(\omega,t)\ \phi(\omega,t)]^T$ be a vector containing the two observed quantities at frequency ω and time t. The KDEs describing the target and interference interaural parameters are, respectively,

$$k_{T\omega}(x') = \sum_t w_T(\omega,t)\,\mathcal{N}(x' \mid x(\omega,t), \Sigma_T(\omega,t)) \tag{2}$$

$$k_{I\omega}(x') = \sum_t w_I(\omega,t)\,\mathcal{N}(x' \mid x(\omega,t), \Sigma_I(\omega,t)) \tag{3}$$

where x' is the arbitrary (ILD,IPD) point at which the estimate is being evaluated, w_T and w_I are the energy of the individual target and interferer observations, and the covariance functions Σ_T and Σ_I are diagonal and are set using Silverman's rule-of-thumb [7, p. 48]. Note that the density is estimated separately for each frequency channel. Note also that both KDEs are formed from the same points, the interaural parameters from the mixture, but weight those points differently using knowledge of the unmixed sources. We use the bounded error complexity reduction of [8] to remove redundant kernels in the estimator while minimally distorting the modeled density. See Figure 1 for an example of target and interferer KDEs and the decision boundary they induce at one particular frequency. The IIPM is then created according to

$$M_{IIPM}(\omega,t) = \begin{cases} 1 & \frac{k_{T\omega}(x(\omega,t))}{k_{I\omega}(x(\omega,t))} \geq \gamma \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where γ is a user-defined threshold. As with the Ideal Binary Mask, we use $\gamma = 1$, meaning that a point in the interaural spectrogram is included if the KDE of the target energy at that pair of interaural parameters is larger than the KDE of the interferer energy.

We compare two different IIPMs in this paper. The first separates signals using the interaural parameters calculated directly from each of the signals involved in a given mixture. We refer to this IIPM as the "training" IIPM. If the bandwidth of the kernels were infinitely small, this IIPM would revert to the Ideal Binary Mask. This can be thought of as over-fitting the density estimator to the particular mixture under analysis, and would have a very low leave-one-out cross-validation likelihood. The training IIPM avoids this by setting the bandwidths according to the rule-of-thumb estimate, which tends to create favorable leave-one-out cross-validations.

The second IIPM uses interaural parameters calculated from different signals, but passed through the same impulse responses as the signals being separated. We refer to this IIPM as the "testing" IIPM. It still uses oracle knowledge of the unmixed signals to construct the KDEs, but to a lesser extent than the "training" IIPM does. It is guaranteed not to over-fit the test data. This IIPM was trained on a set of seven five-digit sentences from the target speaker, a total of 14 seconds of audio, which exhausted our memory resources. The threshold in each band implicitly includes an estimate of the signal-to-noise ratio, so if the per-band SNR between testing and training is significantly different, the threshold should be adjusted. We used similar utterances from the same speaker in the two conditions, so we did not adjust the threshold.

2.2. MESSL

We have made a number of changes to our Model-Based EM Source Separation and Localization (MESSL) algorithm [2] to compare to this upper bound. This algorithm jointly separates and localizes sound sources by probabilistically modeling the interaural parameters of each source according to

$$10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \approx 10^{a(\omega)/20} e^{-j\omega\tau(\omega)} N(\omega,t) \tag{5}$$

The ILD is modeled as a Gaussian with mean $a(\omega)$. The IPD is modeled with a mixture of Gaussians with means $\omega\tau_i(\omega)$, i.e. is constrained to be approximately linear with frequency. Some aspects of phase wrapping are taken into account as well.

As described in [2], MESSL treats the interaural level and phase differences as independent variables, so they only interact through the mask that each generates. Additionally, each point in the spectrogram is assumed to be an independent observation and all points in the spectrogram are treated equally in estimating parameters. Typically, only the IPD parameters are initialized, from a cross-correlation estimate of the source interaural time differences (ITD), $\tau(\omega)$. Other parameters (e.g. ILD parameters) are first estimated from the regions that the initial IPD selects as relevant.

We now describe the modifications that we have made to MESSL to better separate sources in reverberation. The interaural parameters of reverberation differ from those of direct-path signals, even when both originate from the same source. Late reverberation tends to act as a diffuse sound source [9], meaning that its

Table 1: Overall separation results, averaged across all angles. Additions to MESSL are Garbage source, ILD prior, and Reliability.

Algorithm	SNRI ± 95%
IBM	9.55 ± 0.25
IIPM training	7.68 ± 0.18
IIPM testing	6.77 ± 0.14
MESSL+Garb+ILD prior+Rel	5.97 ± 0.20
MESSL+Garb+ILD prior	5.86 ± 0.18
MESSL+Garb	5.39 ± 0.17
MESSL	4.41 ± 0.15

interaural parameters depend only on frequency and microphone separation, and not on the source signal or position. This implies that a single model of interaural parameters can be used for all of the reverberation in a mixture, regardless of the number of sources.

We thus use a single "garbage" model for the reverberation from all of the sources in a mixture. For example, in a mixture of two sources, we instantiate three source models, two for the direct-path components of the sources and one for the reverberation of both sources. The garbage model absorbs points that are poorly described by the direct-path models, keeping them out of both parameter estimates and separation masks.

The garbage model uses the same distributions for ILD and IPD as the source models, but is initialized so that its ILD is a Gaussian with frequency-independent mean of 0 dB and standard deviation of 9 dB. These values were observed to fit the reverberation well in the impulse responses used in the experiments. Its IPD is modeled as a mixture of Gaussians weighted so that all IPDs are equally likely. As the parameters are refined, the IPD model grows to favor certain IPDs at certain frequencies, but there is no discernible pattern across frequency, as there is with direct-path signals.

The second modification to MESSL is the addition of a prior on the ILD, which explicitly captures its dependence on location. MESSL models the ILD as a W-dimensional Gaussian with diagonal covariance, so we use the conjugate normal-Wishart distribution as the prior and compute the maximum *a posteriori* ILD parameters [10]. The mean of this prior is a caricature of the ILDs that we observed in our collection. It starts at 0 dB at 0 kHz and increases linearly to 4 kHz, remaining constant at higher frequencies. Its value, above 4 kHz is proportional to the ITD of the source, with a maximum value of 15 dB when the source is 90° to the right.

This prior distribution acts like a count of virtual observations that are factored into the parameter estimation. The variance of this prior Gaussian determines the number of virtual observations. We use a number of virtual observations that is proportional to signal duration, meaning that the relative strength of the prior is preserved regardless of the number of points that are observed. The actual value was found through testing on a separate dataset to be 4 virtual observations per 100 spectrogram frames.

The final modification to MESSL is the notion of reliability of time-frequency points for parameter estimation. Reverberation diminishes the interaural level difference in proportion to the ratio of direct to reverberant energy [11]. Thus the true ILD parameters for a source (those that would be measured if there were no reverberation) are best estimated from spectral regions with high DRRs.

Because it is an EM algorithm, parameters in MESSL are estimated using weighted sums of sufficient statistics, where points are weighted by their posterior probability under each source model. The prominence of reliable points in the sum was increased by multiplying each posterior by a corresponding reliability weight. Points with a DRR above 0 dB were given a weight of 0.99, and those below were given 0.01. The DRR was calculated from the pre-mixed sources. By cheating, we can determine whether it would be worthwhile to include estimates of reliability from e.g. precedence-effect inspired monaural features [12] or the coherence of the signals at the two ears [13].

Table 2: Regression results. "Standardized" coefficients apply to unit-variance predictors, "original" apply to the predictors in their original units. Uncertainty intervals are for 95% confidence.

Predictor	Unit	Standardized	Original
Initial SNR	dB	-0.57 ± 0.05	-0.24 ± 0.02
Garbage src	binary	0.44 ± 0.05	0.87 ± 0.09
ILD prior	binary	0.24 ± 0.05	0.48 ± 0.09
Reliability	binary	0.16 ± 0.05	0.33 ± 0.09
cos(Angle)	—	-0.14 ± 0.05	-0.33 ± 0.12

3. EXPERIMENTS

We examine the separation performance of the various algorithms as a function of the angle between two sources. The data used in the experiment come from [14]. Binaural room impulse responses were recorded with a KEMAR dummy head in the center of a classroom with a reverberation time (T_{60}) of 565 ms. The source was 1.2 m away from the head. One impulse response was recorded every 15° from $-90°$ to 90°. The target speaker was always located at 0° and the interfering speaker was located at ±15°, ±45°, or ±90°.

The target speech consisted of strings of five digits spoken by a male with natural intonation. The interfering speech is spoken by the same male speaker and consists of read sentences from the TIMIT corpus. At each location the same 32 pairs of target and interferer utterances were used. These pairs were randomly selected from 100 digit strings and 55 sentences.

Results from seven systems are reported. The IBM provides an overall ceiling on performance. The IIPM "training" and "testing" variants described in Section 2.1 provide a ceiling on interaural parameter-based separation. MESSL was evaluated with all eight combinations of garbage source, ILD prior, and reliability, but we only report the results of the four most interesting combinations.

3.1. Evaluation

We measure the performance of the various separations using the signal-to-noise ratio improvement (SNRI), which we have used in our previous work [2]. In it, energy from the direct-path of the target utterance that survives the mask is considered "signal" while energy from the target reverberation or any part of the interferer utterance that survives the mask is considered "noise" along with the target direct-path energy that is eliminated by the mask. Because the mixtures are simulated, we have access to the original signals and can manipulate the impulse responses. By dividing each impulse response at 9.6 ms into an initial impulse and subsequent reverberation, we are able to create separate versions of the direct-path and reverberation of each source.

To compare their relative importance, we ran a linear regression on the SNRI results for all eight combinations of the three

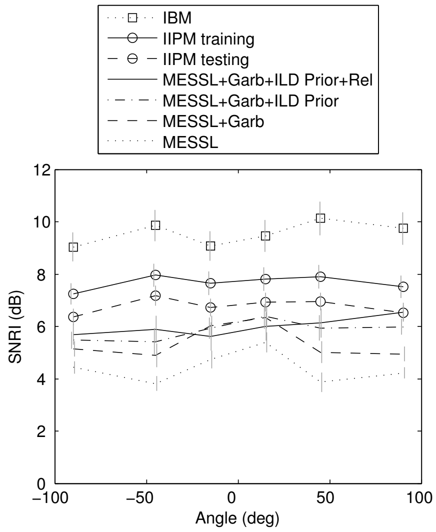


Figure 2: Signal to noise ratio improvement vs separation angle for MESSL variants, the Ideal Binary Mask, and the Ideal Interaural Parameter Masks. Error bars show 95% confidence intervals.

additions to MESSL. The predictors in this regression were various characteristics of the algorithms used and the mixtures that they separated. Specifically, they were the initial SNR of the mixture, use of the garbage source, use of the ILD prior, use of the reliability measure, and the angle between the target and interferer.

The regression was performed on both standardized and the original versions of these predictors. The coefficients for the standardized predictors show the relative importance of the predictors. The coefficients for the original predictors show the partial derivative of the estimate with respect to each predictor, i.e. the increase in predicted SNRI due to one unit increase in each predictor.

4. RESULTS

Table 1 shows the value of defining and comparing to the IIPM. If we considered only the IBM, we would conclude that the enhancements to MESSL, not including the reliability, achieve no more than 28% (1.45 dB) of the 5.14 dB apparent margin for improvement. But within the class of separation systems mediated solely by the interaural parameters, it is the IIPM that should be considered the upper bound, and these improvements have recovered 61% of the margin, taking MESSL within 0.91 dB of the ceiling.

Figure 2 shows the SNRI scores for the various algorithms. The bottom four lines are the four notable variants of MESSL, note that all of the new features increase performance. The top three lines are the ideal binary mask, and the training and testing IIPM. Note the difference in performance between the testing and training IIPM is approximately 1 dB, indicating that the training IIPM is still over-fitting the data to some extent.

Table 2 shows the results of the regression analysis. The overall regression has an R^2-statistic of 0.41, and a residual 95% confidence interval of 1.90 dB. The standardized coefficients show that the greatest influence on the SNR improvement is the initial SNR. Because it is negative, this coefficient implies that a large initial SNR leaves less room for SNR improvement. The coefficient for the original version of that predictor implies that for every dB of initial SNR, the SNR improvement decreases by 0.24 dB. Because the predictors for the three model additions are binary, their coefficients for their original predictors indicate their respective contributions to the SNRI when they are present. There is a small, but significant effect of the angle of separation on MESSL separation performance, a result that must be attributed to the impulse responses used.

5. CONCLUSION

We have introduced the Ideal Interaural Parameter Mask as an upper bound on the class of source separation systems mediated solely by the interaural parameters. After making certain improvements to our MESSL algorithm, its performance approaches the IIPM upper bound to within 0.9 dB. These additions battle the effects of reverberation in various ways. The difference between the performance of the IIPM and the ideal binary mask indicates that further underdetermined, reverberant source separation improvements will probably require exploiting properties of the signals themselves.

6. ACKNOWLEDGMENTS

The authors would like to thank B. Shinn-Cunningham and S. Bressler for the use of their recordings and experimental data.

7. REFERENCES

[1] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Tr. SP*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] M. I. Mandel and D. P. W. Ellis, “EM localization and separation using interaural level and phase cues,” in *WASPAA*, 2007, pp. 275–278.

[3] Y. Li and D. Wang, “On the optimality of ideal binary time-frequency masks,” *Speech Comm.*, vol. 51, no. 3, pp. 230–239, Mar. 2009.

[4] N. Roman, D. Wang, and G. J. Brown, “A classification-based cocktail party processor,” in *NIPS*, 2003, pp. 1425–1432.

[5] S. Harding, J. Barker, and G. J. Brown, “Mask estimation for missing data speech recognition based on statistics of binaural interaction,” *IEEE Tr. ASLP*, vol. 14, no. 1, pp. 58–67, 2006.

[6] E. Parzen, “On estimation of a probability density function and mode,” *An. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.

[7] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, April 1986.

[8] A. T. Ihler, J. W. Fisher III, and A. S. Willsky, “Using sample-based representations under communications constraints,” MIT, Laboratory for Information and Decision Systems, Tech. Rep. 2601, 2004.

[9] Y. Izumi, N. Ono, and S. Sagayama, “Sparseness-based two-channel BSS using the EM algorithm in reverberant environment,” in *WASPAA*, 2007, pp. 147–150.

[10] J. L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE Tr. SAP*, vol. 2, no. 2, pp. 291–298, 1994.

[11] B. Shinn-Cunningham, N. Kopco, and T. J. Martin, “Localizing nearby sound sources in a classroom: Binaural room impulse responses,” *JASA*, vol. 117, no. 5, pp. 3100–3115, 2005.

[12] K. W. Wilson and T. Darrell, “Learning a precedence effect-like weighting function for the generalized cross-correlation framework,” *IEEE Tr. ASLP*, vol. 14, no. 6, pp. 2156–2164, 2006.

[13] C. Faller and J. Merimaa, “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,” *JASA*, vol. 116, no. 5, pp. 3075–3089, 2004.

[14] S. Bressler and B. S. Cunningham, “Effects of pitch and spatial separation on selective attention in anechoic and reverberant environments,” *JASA*, vol. 123, no. 5, p. 2978, 2008.