

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 source separation algorithms that are based purely on the differences between two channels. With three 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 forcing the ILD to be larger than it might otherwise be. The third addition was an oracle reliability measure, 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 where there are more sources than microphones. Mathematically, they are difficult to separate, but one promising approach to separating 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, but overcoming these limitations is still an open problem.

The Ideal Binary Mask, 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 because it uses knowledge of the pre-mixed signals to create the optimal mask in terms of signal-to-noise ratio of the separation. 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 as a single unit. 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 techniques, 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 ILD and ITD 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. Three different oracle algorithms were compared in [6] for three different types of source separation problems: instantaneous mixtures, single-channel time-frequency masking, and multi-channel time-frequency masking.

2. SOURCE SEPARATION SYSTEMS

In this paper, we compare two different source separation systems, the IIPM and MESSL. We now describe these two algorithms 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 parameterization 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, we model them with kernel density estimators (KDEs) [7], nonparametric models similar to histograms.

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

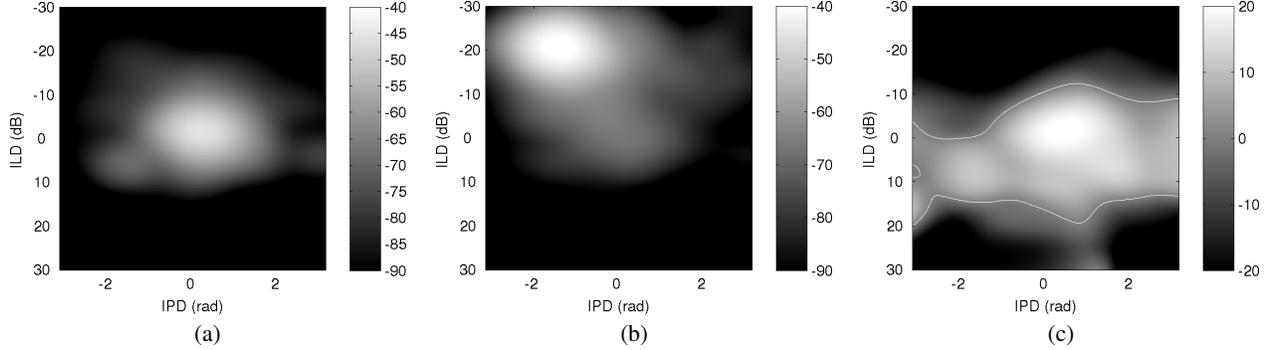


Figure 1: Example kernel density estimates in dB of (a) target energy, (b) interferer energy, (c) energy ratio between target and interferer, the line indicates the decision boundary at 0 dB. From the 4125 Hz band for the target at 0° and interferer at 90° .

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 $x(\omega, t) = [\alpha(\omega, t) \phi(\omega, t)]^T$ be the interaural level and phase differences of the mixture at frequency ω and time t . The KDEs describing the target and interference interaural parameters are, respectively,

$$k_{T\omega}(x') = \sum_t w_T(\omega, t) \mathcal{N}(x' | x(\omega, t), \Sigma_T(\omega, t)) \quad (1)$$

$$k_{I\omega}(x') = \sum_t w_I(\omega, t) \mathcal{N}(x' | x(\omega, t), \Sigma_I(\omega, t)) \quad (2)$$

where 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 [8, p. 48]. Note 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 [9] 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

$$M_{IIPM}(\omega, t) = \begin{cases} 1 & \frac{k_{T\omega}(x(\omega, t))}{k_{I\omega}(x(\omega, t))} \geq \gamma \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where γ is a user-defined threshold. To continue the analogy 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 power at that pair of interaural parameters is larger than the KDE of the interferer power.

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. Our measure 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 sources, passed through the same impulse responses as the

sources 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 and it is guaranteed not to over-fit the test data. We trained on a separate 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 is significantly different between the testing and training conditions, the threshold might need to be adjusted. We used similar utterances from the same speaker in the two conditions, so we did not adjust the threshold.

2.2. MESSL

To compare to this upper bound, we have made a number of changes to our Model-Based EM Source Separation and Localization (MESSL) algorithm [2]. This algorithm jointly separates and localizes sound sources by probabilistically modeling the interaural parameters of each source. The ILD, measured in dB, is modeled as a Gaussian. The IPD is modeled with a mixture of Gaussians, where the mean of each Gaussian is constrained to be linear with frequency and some aspects of phase wrapping are taken into account.

As described in [2], MESSL treats the interaural level and phase differences as independent variables. Only the IPD parameters are typically initialized, from a cross-correlation estimate of the source ITDs. Other parameters of each source (e.g. the ILD parameters) are first estimated from the regions that the initial IPD selects as relevant, so the IPD and ILD 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.

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 [10], meaning that its 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, no matter how many direct-path signals are present.

We thus use a single “garbage” source to model the reverberation from all of the sources in a mixture. For example, in a mixture

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

| Algorithm | SNRI \pm 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 |

of two sources, we instantiate three source models, two for the direct-path components of the sources and one for the garbage source capturing the reverberation of both sources. The garbage source absorbs points that are poorly described by the direct-path models, and so preserves the integrity of the models and keeps extraneous points out of the separation masks.

The garbage model uses the same distributions for ILD and IPD as the source models, but is initialized differently. It is initialized so that its ILD is a Gaussian with frequency-independent mean of 0 dB and standard deviation of 9 dB, values that were observed in the impulse responses we 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 models its relationship with 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 [11]. 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 precision of this prior Gaussian, the inverse of the variance, controls the number of virtual observations. We use a precision that is proportional to the number of observed frames, 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 [12]. 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. Each weight is the posterior probability of a given point under a given source model. By multiplying each weight by a corresponding reliability, reliable points are given more weight in the sum. Points with a DRR above 0 dB were given a reliability of 0.99, and those below were given 0.01. The DRR ratio was calculated from the pre-mixed sources. By cheating, we use this measure as an upper bound

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

| Predictor | Unit | Normalized | Unnormalized |
|-------------|--------|------------------|------------------|
| 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 |

on performance to determine whether it would be worthwhile to include estimates of reliability from precedence-effect inspired monaural features [13] or from the coherence of the signals at the two ears [14].

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 [15]. 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 $\pm 15^\circ$, $\pm 45^\circ$, or $\pm 90^\circ$.

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.

In total eight systems were evaluated. The ideal binary mask 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. And MESSL was evaluated with all eight combinations of the garbage source, ILD prior, and reliability were tested, although we only report the results of the four most typical combinations.

3.1. Evaluation

We measure the performance of the various separations with 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 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,

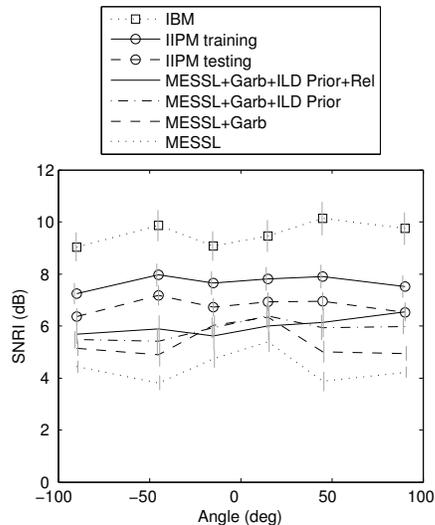


Figure 2: Signal to noise ratio improvement vs separation angle for MESSL variants, the Ideal Binary Mask, and the Ideal Interaural Parameter Masks. Error bars show 95% confidence intervals.

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 normalized and unnormalized versions of these predictors. The coefficients for the normalized predictors show the relative importance of the various predictors on the same scale. The coefficients for the unnormalized predictors show the partial derivative of the estimate with respect to each predictor, i.e. the increase in predicted SNRI due to one unit of increase in each predictor.

4. RESULTS

The value of the IIPM is apparent from Table 1. The enhancements to MESSL (not including the reliability measure) achieve only 28% of the apparent available improvement between its baseline and the ideal binary mask. Over the class of separation systems mediated only by the interaural parameters, our improvements have brought us to within 0.9 dB of the maximum possible performance, that of the IIPM, an increase of 61% over its baseline.

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 normalized 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 normalized version of that coefficient implies that for every dB of initial SNR, the SNR improvement decreases by 0.23 dB. Because the predictors for the three model additions are binary, their unnormalized coefficients 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 results 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 only 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 come from exploiting properties of the signals themselves.

6. ACKNOWLEDGEMENT

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 timefrequency 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. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, Aug. 2007.
- [7] E. Parzen, "On estimation of a probability density function and mode," *An. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.
- [8] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, April 1986.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] A. Ihlefeld and B. Shinn-Cunningham, "Effect of source location and listener location on ild cues in a reverberant room," *JASA*, vol. 115, no. 5, p. 2598, 2004.
- [13] 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.
- [14] W. M. Hartmann, B. Rakerd, and A. Koller, "Binaural coherence in rooms," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 451–462, May 2005.
- [15] 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.