ENFORCING CONSISTENCY IN SPECTRAL MASKS USING MARKOV RANDOM FIELDS

Michael I Mandel[*] *and Nicoleta Roman*[†]

[*] Computer Science and Engineering
The Ohio State University, USA

[†] Computer Science and Engineering
The Ohio State University at Lima, USA

ABSTRACT

Localization-based multichannel source separation algorithms typically operate by clustering or classifying individual time-frequency points based on their spatial characteristics, treating adjacent points as independent observations. The Model-based EM Source Separation and Localization (MESSL) algorithm is one such approach for binaural signals that achieves additional robustness by enforcing consistency in inaural parameters across frequency. This paper incorporates MESSL into a Markov Random Field (MRF) framework in order to additionally enforce consistency in the assignment of neighboring time-frequency units to sources. Approximate inference in the MRF is performed using loopy belief propagation (LBP), and the same approach can be used to smooth any probabilistic source separation mask. The proposed MESSL-MRF algorithm is tested on binaural mixtures of three sources in reverberant conditions and shows significant improvements over the original MESSL algorithm as measured by both signal-to-distortion ratios as well as a speech intelligibility predictor.

Index Terms— Binaural separation, Spectral masking, Markov Random Fields

1. INTRODUCTION

Spectral masking is a technique for suppressing unwanted sound sources in a mixture by applying different attenuations to different time-frequency (T-F) points in a spectrogram [1]. A number of binaural (i.e., two-microphone) systems have combined localization-based clustering with spectral masking (see for example [2–5]). The Model-based Expectation Maximization Source Separation and Localization (MESSL) algorithm performs mask-based separation by clustering T-F units in the joint space of interaural phase differences (IPD), interaural time differences (ITD), and interaural level differences (ILD) [5]. It has also been extended to include frequency-dependent parameters, a prior on the ILD means given the ITDs, and the non-linear warping of probabilities to masks. [6] combined the spatial separation of MESSL with a probabilistic source model. Instead of estimating a single maximum likelihood setting of parameters, [7] uses variational Bayesian inference to estimate posterior distributions over the MESSL parameters. Instead of a grid of ITDs, [8] used random sampling to extract the best IPD-ILD parameters for a multichannel configuration.

One widely recognized problem that arises in mask-based separation is musical noise due to isolated false positive T-F points in the mask. Several approaches have attempted to alleviate this problem by applying a separate smoothing process after estimating their masks [9–12]. In contrast, the current paper proposes embedding the MESSL algorithm into a grid-shaped pairwise Markov random field (MRF) to simultaneously estimate model parameters and smooth T-F masks. In image segmentation applications, these models have been shown to be effective at combining evidence across neighboring pixels. While exact inference in these models is intractable, a number of approximation methods have been shown to be effective, including graph-cuts and loopy belief propagation (LBP) [13]. In addition, learning the parameters of an MRF model is typically intractable, but it has been shown that approximate learning using expectation maximization can provide a reasonable approximation in practice for segmenting noisy images [14, 15].

MRFs have been used in several speech separation systems recently [16, 17]. [17] used an MRF in single-channel audio source separation to smooth masks estimated by non-negative matrix factorization. In an earlier paper [18], the same authors utilized an MRF for binaural audio source separation, but only use interaural level difference as the local feature. They show that utilizing the connections between neighboring time-frequency points can increase signal-to-distortion ratios. Our model differs from [18] in that we are able to utilize IPD and ITD in addition to ILD to perform source separation. We have shown in the past that combining these features performs substantially better than ILD alone [5]. In addition, the current paper describes how the sum-product version of LBP can be used to more efficiently compute smooth probabilistic masks than the Gibbs sampling used by [18].

2. MESSL

Model-based EM source separation and localization (MESSL) [5] is a binaural source separation system that clusters time-frequency points based on similarities in interaural phase and level differences (IPD and ILD). It performs this clustering using expectation maximization (EM) on a Gaussian mixture

model that has been augmented with an additional hidden variable connecting IPD to interaural time difference (ITD). This hidden variable allows it to overcome the maximum frequency limit set by spatial aliasing that affects similar algorithms (e.g., [3]). Avoiding this limitation is especially useful when analyzing binaural recordings from real or dummy heads, where spatial aliasing begins at 1–2 kHz.

The EM clustering results in a probability of each T-F point of the spectrogram being in each source. The expected complex value at each T-F point for each source can then be computed by simply multiplying these probabilities with the observations, corresponding to using the probabilities directly as a soft time-frequency mask (MESSL-P). Alternatively, the probabilities can be transformed into more of a Wiener-like mask using a point-wise sigmoidal non-linearity, with parameters learned on a separate dataset (MESSL-W) [19, §5.2.4]. This point-wise non-linearity empirically improves separation performance, but does not have a clear theoretical justification. We compare both of these approaches to the proposed approach in the experiments and show that the proposed approach without this point-wise non-linearity performs better separation than both of them while having a much clearer theoretical justification.

We now describe the MESSL model. In the absence of additive noise or multiple talkers, which will be considered shortly, a single source, $s(t)$, arriving at two ears, $\ell(t)$ and $r(t)$, through channels $h_\ell(t)$ and $h_r(t)$, can be written as

$$\ell(t) = s(t) * h_\ell(t) \qquad r(t) = s(t) * h_r(t). \tag{1}$$

The ratio of the short-time Fourier transforms, $\mathcal{F}\{\cdot\}$, of both equations is the interaural spectrogram, which provides the observations upon which MESSL operates

$$\frac{\mathcal{F}\{\ell(t)\}}{\mathcal{F}\{r(t)\}} = \frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{2}$$

If the channel to each ear includes a sufficient amount of energy from the direct-path, then these observations can be written as

$$\approx 10^{a(\omega)/20} e^{-j\omega\tau(\omega)} N(\omega,t) \tag{3}$$

where $a(\omega)$ is the ILD (measured in dB), $\tau(\omega)$ is the ITD, and $N(\omega,t)$ is a complex valued, time- and frequency-dependent noise of relatively small magnitude.

In order to avoid spatial aliasing issues, we treat the interaural delay as a discrete hidden variable. Instead of computing $\tau(\omega)$ directly from the $\phi(\omega,t)$ observations, which leads to ambiguous results for higher frequencies, we compare the observations to what we would expect them to be for each of a set of discrete τs, providing unambiguous estimates of the relative support of the observation for each hypothesis. To measure the difference between the IPD predicted by a particular τ and the observed IPD, we define the phase residual $\hat{\phi}$ as

$$\hat{\phi}(\omega,t;\tau) = \arg\left(e^{j\phi(\omega,t)} e^{-j\omega\tau}\right) \tag{4}$$

which is always in the interval $(-\pi,\pi]$. We model this phase residual with a single Gaussian with a per-delay frequency-dependent mean, $\xi_\tau(\omega)$, and standard deviation, $\sigma_\tau(\omega)$,

$$p(\phi(\omega,t) \,|\, \tau,\Theta) = \mathcal{N}\big(\hat{\phi}(\omega,t;\tau) \,|\, \xi_\tau(\omega), \sigma_\tau^2(\omega)\big) \tag{5}$$

which is a good approximation when $\xi_\tau(\omega)$ and $\sigma_\tau(\omega)$ are small relative to π. The ILD is also well modeled by a single Gaussian with frequency-dependent mean and variance

$$p(\alpha(\omega,t) \,|\, \Theta) = \mathcal{N}\big(\alpha(\omega,t) \,|\, \mu(\omega), \eta^2(\omega)\big). \tag{6}$$

The IPD residual and ILD models are assumed to be conditionally independent,

$$\begin{aligned} p(x(\omega,t) \,|\, \tau,\Theta) = \mathcal{N}\big(\hat{\phi}(\omega,t;\tau) \,|\, \xi(\omega), \sigma^2(\omega)\big) \\ \cdot \mathcal{N}\big(\alpha(\omega,t) \,|\, \mu(\omega), \eta^2(\omega)\big), \end{aligned} \tag{7}$$

where $x(\omega,t)$ is the pair $[\phi(\omega,t), \alpha(\omega,t)]$ and Θ represents all of the model parameters. Note that this assumption of independence applies only between the IPD residual and ILD, so does not contradict the well known correlation between ILD and ITD in actual head-related transfer functions, which is enforced instead between the means of these Gaussians, $\xi_\tau(\omega)$ and $\mu(\omega)$. In this work, we model this correlation by enforcing a prior on the ILD based on the initialization of the ITD [19, §5.1]. Because the ILD is modeled as a Gaussian, we use the normal-Wishart distribution, its conjugate, as the prior [20]. The mean of this normal-Wishart is set from the initial ITD using data on their relationship collected from synthetic impulse responses, but the system is not especially sensitive to particular values.

To model multiple sources, we introduce hidden binary random variables $z_{k\tau}(\omega,t)$, which are 1 when time-frequency point ω,t comes from source k and delay τ. Thus our model becomes a Gaussian mixture model, with one Gaussian per combination of source and delay. We can thus use the EM algorithm to learn the maximum likelihood model parameters, Θ, from the observations in an unsupervised fashion, while simultaneously estimating the expected values of the $z_{k\tau}(\omega,t)$. By marginalizing those expected values over delay, we can compute the probability of each time-frequency point coming from each source, probabilistically separating them.

Specifically, in the E step, we compute the likelihood

$$\tilde{\nu}_{k\tau}(\omega,t) = \pi_{k\tau} \cdot p(x(\omega,t) \,|\, \tau,\Theta) \tag{8}$$

where $\pi_{k\tau}$ is the mixing coefficient for the Gaussian indexed by k and τ. We then compute the posterior as

$$\nu_{k\tau}(\omega,t) = \frac{\tilde{\nu}_{k\tau}(\omega,t)}{\sum_{k'\tau'} \tilde{\nu}_{k'\tau'}(\omega,t)}. \tag{9}$$

In the M step, the parameters are updated to be weighted means of sufficient statistics using this posterior [5].

3. PAIRWISE MARKOV RANDOM FIELDS

A Markov random field (MRF) is an undirected graphical model, representing the joint probability of several random variables as a product of potential functions over subsets of those variables [21]. Depending on the structure of the graph, certain quantities can be estimated much more efficiently because of this factorization. This work focuses on pairwise MRFs, in which only pairwise interactions between variables are non-zero and thus only pairwise potential functions are necessary. The joint distribution of random variables $z_1, z_2, \ldots, z_N$ can then be written as

$$p(z_1, z_2, \ldots, z_N) = \frac{1}{Z} \prod_{ij} \psi_{ij}(z_i, z_j) \prod_i \psi_i(z_i) \qquad (10)$$

where $\psi_i(z_i)$ is the potential function on variable z_i by itself, perhaps induced by a corresponding observation, and $\psi_{ij}(z_i, z_j)$ is the pairwise potential function between z_i and z_j, representing compatibilities between their various configurations. Using the sum-product variant of the belief propagation algorithm [22] it is possible to estimate the distribution of each individual variable when all of the others are marginalized away. In the case of tree-structured graphs, belief propagation can compute these quantities exactly. In the case of graphs with loops, it can only approximate these quantities, but it has been shown that such approximations perform well [23].

A message $m_{ij}(z_j)$ represents node i's belief in what the distribution of node j should be. It is computed from the messages incoming to node i by

$$m_{ij}(z_j) \leftarrow \alpha \sum_{z_i} \psi_{ij}(z_i, z_j) \psi_i(z_i) \prod_{\ell \in N(i) \backslash j} m_{\ell i}(z_i) \qquad (11)$$

where α is an arbitrary scaling factor and $N(i)\backslash j$ are the neighbors of i in the graph except for node j. After iterating the passing of these messages until convergence, the algorithm's "belief" in the marginal distribution of z_i can be computed as

$$b_i(z_i) = \alpha \psi_i(z_i) \prod_{\ell \in N(i)} m_{\ell i}(z_i). \qquad (12)$$

4. MESSL-MRF

We propose smoothing MESSL masks by using the MESSL likelihood as the local potential in a grid-shaped pairwise MRF. In the context of such a model, z_i is the random variable representing the source number responsible for the majority of the energy at time-frequency point i[1]. If there are K sound sources, then z_i is a discrete K-dimensional multinomial random variable. In our experiments, K was 3. The grid-shaped MRF then has potentials between every T-F point and its four direct neighbors in time and frequency. Thus the potential function $\psi_{ij}(z_i, z_j)$ represents the compatibility between source z_i dominating T-F point i and source z_j dominating T-F point j. We set the compatibility potentials, $\psi_{ij}(z_i, z_j)$, to

$$\psi_{ij}(z_i, z_j) = \exp(-\beta \delta(z_i, z_j)) \qquad (13)$$

where $\delta(z_i, z_j)$ is the discrete Dirac delta function, which is 1 when $z_i = z_j$ and 0 otherwise, and β is a parameter that we tuned on a separate validation dataset. While simple, this potential is standard in MRF approaches to image segmentation.

More sophisticated compatibility potentials are possible and can be learned from training data. In particular, at low frequencies, ground truth masks tend to be more correlated across time because of the presence of strong lower harmonics. At high frequencies, they are more correlated across frequency because wideband bursts and frication noise. Thus a frequency-dependent compatibility potential could be useful, but we leave this approach for future work.

In MESSL-MRF, the local potential is defined as

$$\psi_{\omega,t}(z_{\omega,t}) = \sum_{\tau} \nu_{k\tau}(\omega, t) \qquad (14)$$

where we have changed notation from indexing hidden variables by i to ω, t and $\nu_{k\tau}(\omega, t)$ is defined in (9). We find the maximum likelihood parameters Θ from the test data using the EM algorithm [14, 15, 24]. Finding the true maximum likelihood parameters in a grid MRF is infeasible for large graphs because the normalization term $1/Z$ couples all of the parameters of the distribution. One can approximate the parameters that maximize this true likelihood by instead finding the parameters that maximize the unnormalized distribution. When using this approximation, the MRF belief propagation step is simply inserted between the E and M steps of a standard EM algorithm. In MESSL, it thus becomes a mask smoothing step. MESSL's E step computes $\nu_{k\tau}(\omega, t)$, which defines the local potential $\psi_{\omega,t}(z_{\omega,t})$ in (14). From these, LBP is run until convergence to compute the soft beliefs, $b_{\omega,t}(z_{\omega,t})$ from (12). These are used to compute updated posteriors

$$\bar{\nu}_{i\tau}(\omega, t) = \nu_{k\tau}(\omega, t) \frac{b_{\omega,t}(z_{\omega,t})}{\sum_{\tau'} \nu_{i\tau'}(\omega, t)}. \qquad (15)$$

And these are used in the standard MESSL M-step updates.

5. EXPERIMENTS

We compared the effectiveness of the proposed MESSL-MRF algorithm with both MESSL-P and MESSL-W on a two-channel separation task involving three simultaneous talkers in reverberation. The speech came from the TIMIT dataset [25]. The impulse responses were simulated using the shoebox room model [26] incorporating measured anechoic HRTFs for the direction of each arriving impulse. The simulated room was $9 \times 5 \times 3.5$ meters, with the microphone approximately at the

[1] For the purposes of the MESSL-MRF discussion, indices i and j are shorthand for T-F coordinates (ω_i, t_i) and (ω_j, t_j).

center of the room at position $(4.47, 2.53, 1.44)$. The sources were positioned 1 meter away, with one source always straight ahead and the other two sources positioned symmetrically on either side at angles between 15° and 90° spaced every 15°. Five different impulse responses were simulated at each angle, with the simulated speaker position moved slightly between each one. At each angle, 40 mixtures were created involving 66 unique, randomly selected TIMIT utterances. Several algorithmic parameters were tuned on another separate set of mixtures using different TIMIT sentences, but the same set of impulse responses.

For all variants of MESSL we use the following settings, based on [5]. For the first 4 EM iterations of 16 total, the Gaussian means and variance for both ILD and IPD are frequency independent. Gradually, the frequency regions across which the means are tied contract until the parameters are fully frequency-dependent. For MESSL-MRF, we increase β over the course of the EM iterations as well, setting it to 0 for 4 iterations (effectively turning off mask smoothing), 0.2 for 4 iterations, and 0.5 for the last 8 iterations. The ITD parameters $\pi_{i\tau}$ are initialized using the localization results from the PHAT-histogram algorithm [27] and the ILD and IPD parameters are initialized to be uninformative. The models include an additional garbage source, which is initialized with flat ILD, ITD, and IPD distributions, and used to absorb diffuse noises, mainly reverberation from the sources. The probability of being in each source is kept constant over time by renormalizing the $\pi_{i\tau}$ parameters appropriately.

The resynthesized time-domain separations were evaluated using the Short Term Objective Intelligibility (STOI) measure [28], and the performance measures provided by the BSS_EVAL toolbox [29]. STOI is designed to predict the intelligibility of source separation results, and provides an output between 0, meaning unintelligible, and 1, meaning perfectly intelligible. The BSS_EVAL toolbox computes the signal-to-distortion ratio (SDR), which is a combination of the signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR). These metrics are measured in dB, with higher values indicating better noise suppression performance. These metrics all require a reference version of the clean signal, and are usually used for analyzing monaural systems. For our binaural task we used the direct-path signals at the appropriate ear as the reference. The separated sources were matched to the appropriate reference source based on both of their locations.

Results from these experiments are shown in Table 1. These results show each metric for each system at each distracter angle averaged across both ears, all three sources, and the 40 mixtures at each angle. It can be seen that MESSL-MRF performs best in all metrics at all angles, improving SDR by approximately 0.6 dB and STOI by 0.02 at all angles. A Bonferroni *post hoc* test with an α of 0.05 conducted on the SDR metric indicates that the MESSL-MRF means are significantly different than the MESSL-P and MESSL-W means at all angles except 15° and 60°. In addition, for the STOI metric, the MESSL-MRF means are significantly different than the MESSL-P and MESSL-W means at all angles except 15° and 75°. For both SDR and STOI, MESSL-P and MESSL-W do not differ significantly at any angle. Thus, incorporating MESSL as the local potential of an MRF does lead to smoother masks and produces separations with both better objective quality and better objective intelligibility. Informal listening tests confirm that there is less noise leakage between sources and less musical noise when using the MRF.

Table 1. Objective metrics averaged across both ears for all three sources in 40 mixtures at each distracter angle. SDR, SIR, and SAR come from the BSS_EVAL toolkit. MESSL-P uses posteriors $\nu_{k\tau}(\omega, t)$ directly as masks, MESSL-W warps them with a point-wise nonlinearity, and MESSL-MRF is the proposed MRF system. Bold entries are significantly better than alternatives for a given angle.

		Distracter angle					
Metric	MESSL	15°	30°	45°	60°	75°	90°
SDR	P	0.04	2.29	3.55	2.28	2.60	2.24
	W	−0.18	2.57	3.75	2.34	2.51	2.44
	MRF	0.71	**3.18**	**4.29**	2.90	**3.26**	**3.17**
STOI	Un-proc	0.55	0.55	0.55	0.55	0.54	0.54
	P	0.64	0.69	0.71	0.69	0.69	0.69
	W	0.62	0.69	0.72	0.69	0.69	0.69
	MRF	0.62	**0.71**	**0.74**	**0.71**	0.71	**0.71**

6. CONCLUSIONS

This paper has introduced a Markov random field formulation for binaural source separation. It uses the MESSL model as the local potential and finds assignments of time-frequency points to sources so that neighboring points tend to be assigned to the same source. It uses loopy belief propagation for approximate inference in this model and expectation maximization for approximate learning. This assignment has the effect of smoothing the estimated mask, reducing musical noise and noise leakage. Future work includes learning a frequency-dependent pairwise potential function and utilizing the max-sum formulation of LBP to convert soft masks into globally consistent maximum *a posteriori* binary masks.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1409431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation*

by Humans and Machines, Divenyi P., Ed., pp. 181–197. Kluwer, 2005.

[2] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acous. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[3] Ö Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Tr. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Sig. Proc.*, vol. 87, pp. 1833–1847, 2007.

[5] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Tr. ASLP*, vol. 18, no. 2, pp. 382–394, 2010.

[6] R. Weiss, M. I. Mandel, and D. W. P. Ellis, "Combining localization cues and source model constraints for binaural source separation," *Sp. Comm.*, vol. 53, no. 5, pp. 606–621, 2011.

[7] A. Deleforge, F. Forbes, and R. Horaud, "Variational EM for binaural sound-source separation and localization," in *Proc. ICASSP*, 2013, pp. 76–79.

[8] J. Traa, M. Kim, and P. Smaragdis, "Phase and level difference fusion for robust multichannel source separation," in *Proc. ICASSP*, 2014, pp. 6687–6690.

[9] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. ICASSP*, 2008, pp. 45–48.

[10] E. Grais and H. Erdogan, "Spectro-temporal post-smoothing in NMF based single-channel source separation," 2012, pp. 584–588.

[11] E. Vincent, "An experimental evaluation of wiener filter smoothing techniques applied to under-determined audio source separation," in *Research Report RR-7261*, 2010.

[12] M. Cobos and J. Lopez, "Maximum a posteriori binary mask estimation for underdetermined source separation using smoothed posteriors," *IEEE Tr. ASLP*, vol. 20, no. 7, pp. 2059–2064, 2012.

[13] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Tr. PAMI*, vol. 30, no. 6, pp. 1068–1080, 2008.

[14] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE Tr. Med. Img.*, vol. 20, no. 1, pp. 45–57, 2001.

[15] D. B. Gu and J.X. Sun, "EM image segmentation algorithm based on an inhomogeneous hidden MRF model," *IEEE Vis., Im. & Sig. Proc.*, vol. 152, no. 2, pp. 184–190, 2004.

[16] S. Liang, W. Liu, and W. Jiang, "Integrating binary mask estimation with MRF priors of cochleagram for speech separation," *IEEE SPL*, vol. 19, no. 10, pp. 627–630, 2012.

[17] M. Kim and P. Smaragdis, "Single channel source separation using smooth nonnegative matrix factorization with markov random fields," in *Proc. MLSP*, 2013, pp. 1–6.

[18] M. Kim, P. Smaragdis, G. G. Ko, and R. A. Rutenbar, "Stereophonic spectrogram segmentation using markov random fields," in *Proc. MLSP*, 2012, pp. 1–6.

[19] Michael I. Mandel, *Binaural Model-Based Source Separation and Localization*, Ph.D. thesis, Columbia University, Feb. 2010.

[20] 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.

[21] J. Besag, "On the statistical analysis of dirty pictures (with discussion)," *J Royal Stat. Soc., B*, vol. 48, no. 3, pp. 259–302, 1986.

[22] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann, 1988.

[23] J. Yedidia, W. Freeman, and Y. Weiss, "Generalized belief propagation," in *Advances in Neural Information Processing Systems*. 2000, pp. 689–695, MIT Press.

[24] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J Royal Stat. Soc., B*, pp. 1–38, 1977.

[25] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403, 1993.

[26] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acous. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[27] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Tr. Sys., Man, Cyb., C*, vol. 32, no. 4, pp. 474–484, 2002.

[28] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Tr. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Tr. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.