INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Tight integration of spatial and spectral features for BSS with Deep Clustering embeddings

Lukas Drude, Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering, Paderborn, Germany
{drude,haeb}@nt.upb.de

Abstract

Recent advances in discriminatively trained mask estimation networks to extract a single source utilizing beamforming techniques demonstrate, that the integration of statistical models and deep neural networks (DNNs) are a promising approach for robust automatic speech recognition (ASR) applications. In this contribution we demonstrate how discriminatively trained embeddings on spectral features can be tightly integrated into statistical model-based source separation to separate and transcribe overlapping speech. Good generalization to unseen spatial configurations is achieved by estimating a statistical model at test time, while still leveraging discriminative training of deep clustering embeddings on a separate training set. We formulate an expectation maximization (EM) algorithm which jointly estimates a model for deep clustering embeddings and complex-valued spatial observations in the short time Fourier transform (STFT) domain at test time. Extensive simulations confirm, that the integrated model outperforms (a) a deep clustering model with a subsequent beamforming step and (b) an EM-based model with a beamforming step alone in terms of signal to distortion ratio (SDR) and perceptually motivated metric (PESQ) gains. ASR results on a reverberated dataset further show, that the aforementioned gains translate to reduced word error rates (WERs) even in reverberant environments.

Index Terms: blind source separation, deep clustering, expectation maximization, beamforming

1. Introduction

Traditionally, multi channel blind speech separation is tackled with statistical model based separation systems. In particular approaches that exploit the sparseness of speech in the STFT domain have become very popular [1, 2, 3, 4, 5, 6]. The majority of these techniques neglects frequency dependencies carrying out separation on each frequency separately, which lead to the permutation problem: Even if the source separation were perfect for each frequency bin, it is likely, that component one of a given frequency bin does not correspond to the same speaker as component one of another frequency bin [7]. Notable exceptions either apply a frequency normalization [8] or estimate statistics which are shared across frequencies [9].

Single channel source separation is an even harder problem than multi channel source separation. Shallow blind decomposition techniques, such as Nonnegative Matrix Factorization, have met only with limited success [10, 11]. Recently, deep neural network based approaches have shown promise. In particular, deep clustering [12] and its variants [13] are a great step forward towards single channel speech separation: A neural network is trained to learn embeddings from the two-dimensional time-frequency representation of the signal, such that embeddings belonging to the same source form clusters. For the computation of the embeddings correlations in the speech signal both in time and frequency direction are exploited. An attractive property of deep clustering is the fact, that the network is not fixed to a predefined number of speakers. In fact, the number of speakers in the mixture may be different in training than in testing and need not be specified beforehand. This stands in contrast to other neural network based separation techniques, e.g., the permutation invariant training [14].

While single channel source separation relies on spectral properties of the speech signal, multi channel statistical model based solutions exploit the spatial diversity of the sources. In this contribution, we show how to integrate the two and exploit both spectral and spatial cues jointly: The spectral properties of the sources, which are captured by the deep clustering embeddings, and the spatial properties, which are represented by the complex observation vector, are jointly exploited in an EM algorithm to estimate time frequency masks for each speaker. That way, we avoid the aforementioned frequency permutation problem of multi channel model based approaches [7], mitigate the need for a careful initialization and overall achieve better separation performance.

The deep clustering model is trained off-line on a separate training set, whereas the EM algorithm to estimate the mixture model parameters and the masks is carried out at test time on an utterance per utterance basis. Thus, the estimation is independent of the number of microphones and the microphone configuration. The time frequency masks are then used to estimate power spectral density (PSD) matrices for each target speaker and each target speaker's interferences. These matrices are then employed to estimate a generalized eigenvalue (GEV) beamformer [15, 16], which is optimal in terms of the expected output SNR gain.

We evaluate the efficacy of both the proposed hybrid EM and the GEV beamformer on a setting as close as possible to the original deep clustering contribution [12, 17]. However, we give up their assumption of an anechoic environment and create somewhat more realistic spatial observations by convolving the clean speech with artificially generated room impulse responses. Although the proposed algorithm can conceptually be extended to noisy data, we leave this for future work.

To thoroughly analyze, where the different gains come from, we first reconstruct the approach by Isik et al. [17]: The single channel deep clustering method is carried out on a reference channel, followed by a subsequent mask refinement network. The resulting mask is then used as a gain function to obtain a clean speech estimate, similar as in earlier works on soft masks [18]. Subsequently, we replace the mask refinement network and the masking step with the GEV beamformer to measure the gains of a multi-channel model. With said beamformer, we then compare different methods for mask estimation: First, we directly use the deep clustering masks, obtained by a k-means algorithm [19] from the deep clustering embeddings. Then we compare with a purely spatial model based

http://dx.doi.org/10.21437/Interspeech.2017-187

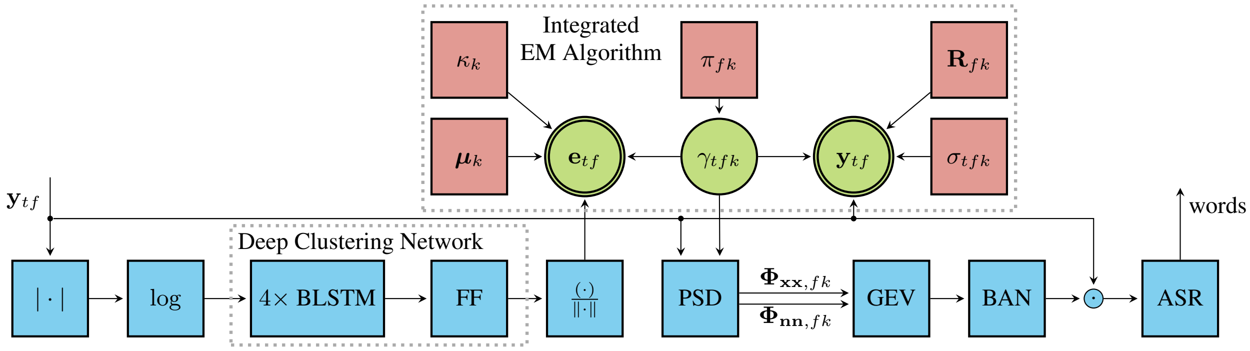

Figure 1: *Integrated EM algorithm with deep clustering embeddings as spectral features and normalized complex observations as spatial features. Blue boxes indicate processing units. Green circles depict random variables, where doubly circled elements are observable random variables, i.e. the embedding vectors $\mathbf{e}_{tf}$ do not change during EM iterations. Red boxes are model parameters which are estimated during test time.*

EM algorithm, which does not use deep clustering. Finally we present results for the integrated mask estimation using deep clustering and spatial models. Results are presented in terms of objective signal quality measures. Further, we provide WER results for selected systems.

2. Signal model

A convolutive mixture of K independent source signals s_{tfk}, captured by D sensors is approximated in the STFT domain:

$$\mathbf{y}_{tf} = \sum_k \mathbf{h}_{fk}\, s_{tfk} + \mathbf{n}_{tf}, \tag{1}$$

where $\mathbf{y}_{tf}$, $\mathbf{h}_{fk}$ and $\mathbf{n}_{tf}$ are the D-dimensional observed signal vector, the unknown acoustic transfer function vector of source k and the vector of noise signals, respectively. Further, t and f specify the time frame index and the frequency bin index, respectively. Since speech signals are sparse in the STFT domain, we may assume that a time frequency slot is occupied either by a single source and noise or by noise only.

3. Deep clustering

Similar to [12, 17] a multi-layered bidirectional long short term memory network (BLSTM) [20] is trained on single channel mixtures to map from the $T \cdot F$ spectral features (log-magnitude spectrum) to the same number of E-dimensional embedding vectors $\mathbf{e}_{tf}$, where $\|\mathbf{e}_{tf}\|_2 = 1$.

The objective during training is to minimize the Frobenius norm of the difference between the estimated and true affinity matrix:

$$J(\theta) = \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_\mathrm{F}^2 = \left\|\mathbf{E}\mathbf{E}^\mathrm{T} - \mathbf{C}\mathbf{C}^\mathrm{T}\right\|_\mathrm{F}^2, \tag{2}$$

where $\hat{\mathbf{A}}$ and $\mathbf{A}$ are the estimated and ground truth affinity matrices. The entries $A_{n,n'}$ encode, whether observation n and n' belong to the same source ($A_{n,n'} = 1$, and zero else). Correspondingly, the embeddings are stacked in a single matrix $\mathbf{E}$ with shape $(TF \times E)$ and the ground truth one-hot vectors describing which time frequency slot belongs to which source are stacked in a single matrix $\mathbf{C}$ with shape $(TF \times K)$, such that $C_{nk} = 1$, if observation n belongs to source k and $C_{nk} = 0$ otherwise.

During training, the network is encouraged to move embeddings belonging to the same source closer together while pushing embeddings which belong to different sources further apart. After training, the embeddings, which are normalized to unit-length, can be clustered to obtain time frequency masks for each source. The original work used k-means clustering.

4. Von-Mises-Fisher Time-Variant Complex-Gaussian Mixture Model

A schematic overview of the proposed system is given in Fig. 1. The complete algorithm is summarized in Alg. 1. In the following, we described the individual components:

In general, an integrated spectral and spatial model is formulated by factorizing the complete data likelihood function:

$$\mathcal{L} = \prod_{tf} p(\mathbf{e}_{tf} | z_{tf} = k; \theta_\text{spectral}) \cdot p(\mathbf{y}_{tf} | z_{tf} = k; \theta_\text{spatial}) \cdot p(z_{tf}; \boldsymbol{\pi}), \tag{3}$$

where $\mathbf{e}_{tf}$ and $\mathbf{y}_{tf}$ are the spectral and spatial features, z_{tf} are the latent random variables and θ_spectral and θ_spatial are the spectral and spatial parameters to be estimated at test time. This decomposition is valid, if the spatial model does not make use of spectral information and vice versa (compare [21]).

Possible spatial models for an integrated EM are, e.g., complex angular-central Gaussian mixture models (cACGMMs) [22], complex Bingham mixture models (cBMMs) [23] and complex Watson mixture models (cWMMs) [3, 4, 24] since they operate on unit-length observations $\tilde{\mathbf{y}}_{tf} = \mathbf{y}_{tf}/\|\mathbf{y}_{tf}\|$.

Another alternative is to use a time-variant complex Gaussian mixture model (TV-cGMM) [9]. For this model, however, the independence assumption of Eq. (4) is, strictly speaking, not valid, because it models the complex observation vectors $\mathbf{y}_{tf}$ directly instead of the orientation vector $\tilde{\mathbf{y}}_{tf}$. However, the power of each time-frequency slot, which is a spectral property, is effectively factored out in each iteration, and therefore the independence assumption is still approximately valid. As shown in the appendix of [22], the EM algorithms of the cACGMM and the TV-cGMM are theoretically equal. Nevertheless, we experienced the implementation of the TV-cGMM to be numerically more stable, specifically in almost silent regions[1].

Consequently, without constraining the integrated model in general, we here use a time-variant (circularly-symmetric and zero mean) complex Gaussian observation model for the spatial features with decoupled variance and correlation matrices in a similar formulation as in [25]:

$$\begin{aligned} p(\mathbf{y}_{tf} | z_{tf} = k; \theta_\text{spatial}) &= \text{TV-cG}(\mathbf{y}_{tf}; \sigma_{tfk}, \mathbf{R}_{fk}) \\ &= \frac{1}{\det(\pi \mathbf{R}_{fk})} \mathrm{e}^{-\mathbf{y}_{tf}^\mathrm{H} \sigma_{tfk}^{-1} \mathbf{R}_{fk}^{-1} \mathbf{y}_{tf}}, \end{aligned} \tag{4}$$

where $\mathbf{R}_{fk}$ is a spatial correlation matrix and σ_{tfk} can be interpreted as a (possibly scaled) local power estimate.

[1] Normalization is re-estimated in each step and spatial covariance matrices are unit-trace normalized.

Since the deep clustering embeddings are real-valued and normalized to unit lengths, a von-Mises-Fisher (vMF) observation model is a suitable choice:

$$p(\mathbf{e}_{tf}|z_{tf}=k;\theta_{\text{spectral}}) = \text{vMF}(\mathbf{e}_{tf};\boldsymbol{\mu}_k,\kappa_k)$$
$$= c_{\text{vMF}}(\kappa_k)\mathrm{e}^{\kappa_k\boldsymbol{\mu}_k^{\mathrm{T}}\mathbf{e}_{tf}}, \qquad (5)$$

where $c_{\text{vMF}}(\kappa_k)$, $\boldsymbol{\mu}_k$ and κ_k are the normalization term [26] and the class-dependent mean and concentration, respectively.

Code for the EM algorithms are available online [2].

4.1. E-step

$$\gamma'_{tfk} = \pi_{fk}\text{vMF}^{\alpha}(\mathbf{e}_{tf};\boldsymbol{\mu}_k,\kappa_k) \cdot \text{TV-cG}^{\beta}(\mathbf{y}_{tf};\sigma_{tfk},\mathbf{R}_{fk}), \qquad (6)$$

$$\gamma_{tfk} = \gamma'_{tfk} \Big/ \sum_k \gamma'_{tfk}, \qquad (7)$$

where π_{fk} are the frequency-dependent class weights and γ_{tfk} are the class affiliation posteriors (masks) for each time-frequency point and each source.

4.2. M-step

$$\pi_{fk} = \pi'_{fk} \Big/ \sum_k \pi'_{fk} \text{ with } \pi'_{fk} = \sum_t \gamma_{tfk}, \qquad (8)$$

$$\boldsymbol{\mu}_k = \frac{\mathbf{r}_k}{\|\mathbf{r}_k\|} \text{ with } \mathbf{r}_k = \sum_{t,f} \gamma_{tfk}\mathbf{e}_{tf}. \qquad (9)$$

Although the concentration parameter κ_k has to be obtained via an implicit equation, an approximation can be formulated in an explicit way, where E is the embedding dimension [26]:

$$\kappa_k = \frac{\bar{r}_k E - \bar{r}_k^3}{1-\bar{r}_k^2} \text{ with } \bar{r}_k = \|\mathbf{r}_k\| \Big/ \sum_{t,f}\gamma_{tfk}. \qquad (10)$$

The parameters of the spatial observation model can be updated as follows[3] [25]:

$$\sigma_{tfk} = \frac{1}{D}\operatorname{tr}\left(\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathrm{H}}\mathbf{R}_{fk}^{-1}\right), \qquad (11)$$

$$\mathbf{R}_{fk} = \sum_t \gamma_{tfk}\frac{1}{\sigma_{tfk}}\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathrm{H}} \Big/ \sum_{t,k}\gamma_{tfk}. \qquad (12)$$

5. GEV beamforming

The GEV beamformer has proven to be robust with respect to numerical instabilities and yields high improvements in terms of signal to noise ratio (SNR) gain as well as WER reduction, while often outperforming the frequently used minimum variance distortionless response (MVDR) beamformer [15, 27]. Within this work, we will employ the GEV beamformer as a means to separate concurrent target speakers. The GEV beamforming approach maximizes the expected SNR gain for a given target k at the beamformer output $z_{tfk} = \mathbf{w}_{fk}^{\mathrm{H}}\mathbf{y}_{tf}$:

$$\mathbf{w}_{\text{GEV},fk} = \underset{\mathbf{w}_{fk}}{\operatorname{argmax}} \frac{\mathbf{w}_{fk}^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{xx},fk}\mathbf{w}_{fk}}{\mathbf{w}_{fk}^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{nn},fk}\mathbf{w}_{fk}}, \qquad (13)$$

where $\mathbf{w}_{fk}$ is the beamforming vector (weight vector) and $\boldsymbol{\Phi}_{\mathbf{xx},fk}$ and $\boldsymbol{\Phi}_{\mathbf{nn},fk}$ are the estimated target and noise covariance matrices:

$$\boldsymbol{\Phi}_{\mathbf{xx},fk} = \sum_t \gamma_{tfk}\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathrm{H}}, \; \boldsymbol{\Phi}_{\mathbf{nn},fk} = \sum_{t,k'\neq k}\gamma_{tfk'}\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathrm{H}}. \qquad (14)$$

Algorithm 1 Source separation algorithm for vMF-TV-GMMs. All steps are performed during test time.

```
Calculate deep clustering embeddings $\mathbf{e}_{tf}$.
Initialize affiliations $\gamma_{tfk}$ with k-means clustering on $\mathbf{e}_{tf}$.
while not converged do
    E-step: Obtain masks $\gamma_{tfk}$ with Eqs. (6) – (7).
    M-step: Eqs. (8) – (12).
end while
Calculate PSD matrices based on EM result with Eq. (14).
Obtain GEV vector and BAN filter.
Obtain source estimate with $z_{tfk} = g_{\text{BAN},fk}\mathbf{w}_{\text{GEV},fk}^{\mathrm{H}}\mathbf{y}_{tf}$.
```

Subsequently, a blind analytic normalization (BAN) postfilter is applied to reduce speech distortions [15, 16].

6. Evaluation

6.1. Setup details

Single channel utterances from the Wall Street Journal corpus [28] are mixed according to the file lists provided by MERL[4]. The training, cross-validation and test set contain 20000, 5000 and 3000 mixtures, respectively. Training and testing is only performed on mixtures with two sources, although the deep clustering framework allows more speakers as well. Since the source signals have different lengths, the images were cut to the minimum lengths of both. The audio files are downsampled to 8 kHz and an STFT (size: 512, shift: 128) with a Blackman window is applied.

We employed a deep clustering model with a similar architecture as in [17]. We used four BLSTM layers and a single fully-connected layer, where each BLSTM layer consists of 300 forward and 300 backward units and the feed forward network consists of 257 units corresponding to the frequency bins [20].

The model is trained with stochastic gradient decent using the ADAM [29] optimization scheme on the log amplitude spectrum of the clean features and uses ideal binary masks as targets. Only the time frequency slots containing 98 % of the total energy were considered as bins occupied by speech. The gradients of all other slots are zero. This leads to a faster convergence and better results, since the model does not waste capacity on slots, which do not belong to any of the target speakers.

To further improve the recipe, dropout with $p = 0.5$ dropout rate was used in the forward connections [30]. In contrast to [17] we did not observe any further gains neither with recurrent dropout nor weight decay. Instead of curriculum learning as in [17], we observed best results when training on entire utterances for 100 epochs with a learning rate of 10^{-3} and another 100 epochs with a learning rate of 10^{-4}.

Using sequence normalization [31] (batch normalization [32], where statistics are obtained in time direction instead of batch direction) also greatly enhanced convergence speed.

The test utterances are reverberated using the Image Method [33] with different random reverberation times (see Figs. 2 and 3). To do so, we randomly sample six microphone positions (approximately on a circle with $r = 20\,\text{cm}$), two speaker positions between 1 m and 2 m away from the array center (no minimum angular distance enforced) and room sizes with approximately 8 m × 6 m × 3 m. Subsequently, the reverberated signals are mixed according to the MERL file lists.

To obtain meaningful word error rates, both hypothesis transcriptions are compared with the reference transcription of

[2] https://github.com/fgnt/dc_integration

[3] In practice, the correlation matrix inverse is obtained via an eigenvalue decomposition. That way, the inverse can be stabilized by clipping the eigenvalue spread to i.e. 10^{10}. Additionally, the determinant can then be calculated as the product of eigenvalues.

[4] http://www.merl.com/demos/deep-clustering

Table 1: *Word error rates for selected models with random reverberation times T_{60} in the range* 50 ms *to* 100 ms.

Model	Extraction	WER / %
k-means + repair	Masking	65.8
k-means	GEV	42.4
vMF-TV-cGMM 0.9	GEV	**29.7**
TV-cGMM	GEV	33.6

the shorter utterance as in [17]. Transcriptions are obtained using the standard Kaldi WSJ recipe (GMM-HMM recognizer as in [17] with LDA and MLLT trained on the `train_si84` dataset (`tri2b` model)) [34]. Please note that the training of the deep clustering network as well as the training of the acoustic model have been carried out on non-reverberant speech.

6.2. Results

Figs. 2 and 3 show the SDR [35] gains and perceptual evaluation of speech quality (PESQ) [36] gains for two different reverberation conditions and a variety of separation methods (described from left to right):

The baseline is the single channel system with an additional repair network (two BLSTM layers, one fully-connected layer) trained to enhance the single channel masks according to [17]. It provides the lowest SNR gains, which is plausible, since it does not make use of more than one channel. In correspondence to Tbl. 1, it leads to the worst WER.

Next, we use the deep clustering model and a simple k-means, as in the original recipe. Additionally, we use the GEV beamformer as in Section 5 on all six channels. As expected, the gains are already much higher than with the original recipe. It comes to no surprise that the gains are lower for the setup with a higher reverberation time.

Now, the k-means clustering is replaced by an EM on the von-Mises-Fisher mixture model (vMFMM). Although the vMF concentration parameter can be updated during EM iterations, it turned out to be more robust to keep it fixed: $\kappa_k = 100$. The performance is almost the same as with the k-means algorithm. This proves, that all further gains can not be attributed to the vMF observation model itself.

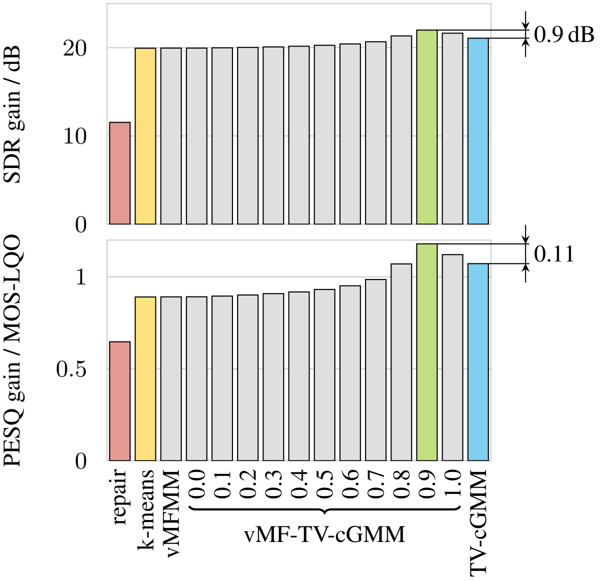

Figure 2: *SDR and PESQ gains for different models with random reverberation times T_{60} in the range* 50 ms *to* 100 ms.

Next, we evaluate the integrated EM algorithm for different exponential weights β for the spatial model, where the exponential weight for the spectral model is $\alpha = 1 - \beta$. As known from integrating language models with acoustic models, it turns out, that an optimal weighting between both models can improve performance. Note, however, that changing the concentration parameter κ of the vMF observation model has the same effect as changing the weights $(\exp(\kappa\boldsymbol{\mu}^{\mathrm{T}}\mathbf{e})^{\alpha} = \exp(\alpha\kappa\boldsymbol{\mu}^{\mathrm{T}}\mathbf{e})$.

Finally, we provide the TV-cGMM results, which are deep clustering agnostic. Therefore, initialization was random and the permutation problem had to be solved additionally. It can be seen that the best results are obtained by the integrated model which uses deep clustering and the spatial model. The best combination achieved an SDR and PESQ gain of 16.9 dB and 0.43, respectively, compared to 14.6 dB and 0.32 for the purely spatial model with random reverberation times T_{60} in the range 200 ms to 300 ms.

Tbl. 1 shows the word error rates for selected models. It can be observed, that the SDR and PESQ gains transfer to WER improvements. The rather high absolute WERs can be attributed to the fact, that the used acoustic model never saw reverberated data and is just trained on the small train dataset (`train_si84`).

7. Conclusions

Summing up, we presented a way to tightly integrate discriminatively trained single-channel spectral models with a statistical model based multi-channel approach. We proved, that this yields notable SDR, PESQ and WER gains in unseen test conditions. We attribute this to the fact, that the statistical model parameters are estimated during test time and that discriminatively trained spectral models nicely transfer knowledge from a training corpus.

8. Acknowledgements

Computational resources were provided by the Paderborn Center for Parallel Computing. We thank S. Watanabe (MERL) for fruitful discussions and insights in the Kaldi recipe.

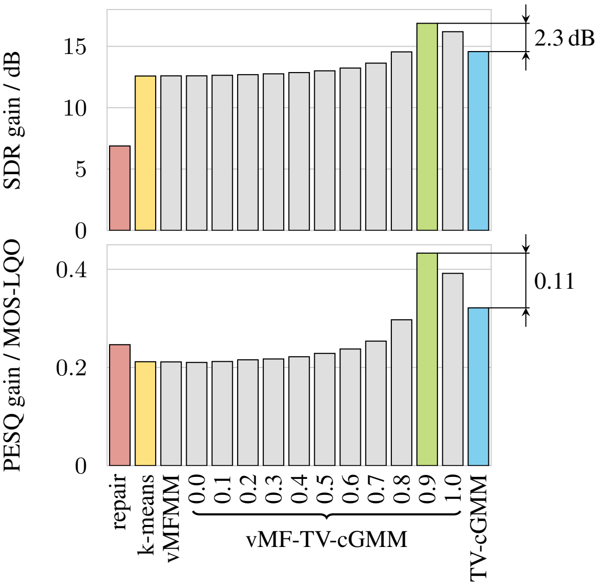

Figure 3: *SDR and PESQ gains for different models with random reverberation times T_{60} in the range* 200 ms *to* 300 ms.

9. References

[1] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

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

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[4] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[5] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[6] L. Drude, C. Boeddeker, and R. Haeb-Umbach, "Blind speech separation based on complex spherical k-mode clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[7] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2007.

[8] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Normalized observation vector clustering approach for sparse source separation," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2006.

[9] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[11] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF – half-baked or well done?" *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015.

[12] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[13] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[14] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[15] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[16] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, 2011.

[17] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech*, 2016.

[18] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[19] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[20] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[21] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[22] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2016.

[23] ——, "Modeling audio directional statistics using a complex Bingham mixture model for blind source extraction from diffuse noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[24] I. Jafari, R. Togneri, and S. Nordholm, "On the use of the Watson mixture model for clustering-based under-determined blind source separation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[25] T. Higuchi, T. Yoshioka, and T. Nakatani, "Optimization of speech enhancement front-end with speech recognition-level criterion," *Interspeech*, 2016.

[26] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *Journal of Machine Learning Research*, vol. 6, no. Sep, pp. 1345–1382, 2005.

[27] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, 2017.

[28] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," *Philadelphia: Linguistic Data Consortium*, 1993.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint:1412.6980*, 2014.

[30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint:1502.03167*, 2015.

[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[35] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.