ON LINEAR AND MIXMAX INTERACTION MODELS FOR SINGLE CHANNEL SOURCE SEPARATION

Robert Peharz and Franz Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology

ABSTRACT

For model-based single channel source separation, one typically assumes a linear interaction model, i.e. that the mixture magnitude spectrogram is the sum of the individual source magnitude spectrograms. In the log-domain, the MIXMAX interaction model is the corresponding approximation for the linear model. Hence, one would expect similar performance for both approaches. However, in this paper we empirically show that this is not the case for vector-quantizer-based (VQ) single channel source separation. We propose factorial linear-VQ, the linear counterpart to factorial max-VQ, and compare the two methods in systematic source separation experiments. Linear-VQ performs significantly better than max-VQ for comparable code-book sizes and behaves more robustly in the presence of additive white noise. Furthermore, we compare resynthesis properties of binary and continuous time-frequency masks. While binary masks achieve a higher interference suppression, the use of continuous masks results in a consistently better signal quality.

Index Terms: source separation, single channel, VQ, time-frequency masking

1. INTRODUCTION

Audio source separation is an important problem in speech and audio processing, with a large number of potential applications, such as preprocessing for automatic speech recognizers in noisy environments, audio post-processing, intelligent hearing-aids, word spotting and audio information retrieval, to name but a few. When sources shall be extracted from a monophonic mixture, we have an instance of the generally ill-posed single channel source separation problem (SCSS). Besides computational auditory scene analysis (CASA) [1], which aims to mimic low-level separation and grouping mechanisms of the biological auditory system, there exists a statistical, model-based approach for SCSS. Generally, in model-based approaches, source specific models are used to infer source signal estimates from the mixture recording. Typically, these models are trained a priori on a time-frequency representation of clean source specific signals. An estimation of the source spectra can then be used to calculate a binary or continuous time-frequency mask for resynthesis. For model-based approaches, the factorial HMM model and the factorial max-VQ model proposed by Roweis [2, 3] can be considered as key-work. Further related models can be found in [4, 5, 6], and in references therein.

Usually, phase information is discarded beforehand, since it is more difficult to handle and considered less important than the magnitude of the time-frequency bins. Some authors prefer to use magnitude spectra, while others work in the log-spectral domain. However, it is not clear whether one of these two approaches is more beneficial than the other. As far as we know, there is no theoretical justification for the use of log-spectra for SCSS, except possibly a biological one, since human beings perceive dynamics on a logarithmic scale. For magnitude spectra, one typically assumes that the mixture spectrum can be approximated by the sum of the source spectra, which is exactly true only for complex-valued spectra. The equivalent approximation for log-magnitude spectra is the $\log \sum \exp$ (softmax) function. Further approximating the softmax with the max function, leads to the well-known MIXMAX approximation, first used by Nadas et al. [7]. In [8], it was shown that the MIXMAX approximation is a nonlinear MMSE estimator for the case of two interfering sources, under the rather mild assumption of uniformly distributed phase differences in each time-frequency bin.

In this paper, we compare the linear and the MIXMAX interaction model by means of the factorial vector quantizer (VQ) model, which was introduced by Roweis for the MIXMAX approach [3]. We propose a correspondent model for (linear) magnitude spectra, and compare the two systems on mixture utterances from the database of Cooke et al. [9]. As a second contribution, we empirically study the advantages and disadvantages of binary and continuous masks, when used for resynthesis. We consider a time-frequency representation of the source signals and the mixture, by transforming the time signals via the short-time Fourier transform (STFT). With $\mathbf{s}^m$, we denote the short-time spectrum of the m^{th} source, and $\mathbf{x}$ denotes the short-time spectrum of the mixture, each containing D frequency bins. The frame index is omitted, since the presented VQ-based methods operate in a frame-wise manner.

The paper is organized as follows. In section 2, we review the factorial max-VQ system [3]. In section 3, we propose its linear counterpart, factorial linear-VQ. In section 4, we describe the resynthesis method using binary or continuous masks. In section 5, we present our experiments and section 6 concludes the paper.

2. FACTORIAL MAX-VQ

In the factorial max-VQ model [3], the spectrum of the m^{th} source is modeled by a VQ with codebook $\mathbf{W}^m = [\mathbf{w}_1^m, \ldots, \mathbf{w}_{K^m}^m]$, containing K^m code-vectors. Each VQ selects its code-vector independently of the others. For the m^{th} source, the k^{th} code-vector is picked with prior probability $\pi_k^m = p(z^m = k)$, $k, z^m \in \{1, \ldots, K^m\}$, where z^m denotes the index of the selected codebook entry. The codebooks $\mathbf{W}^m$ are obtained by k-means training, applied to log-magnitude spectrograms of clean, source specific training data. The priors $\boldsymbol{\pi}^m = (\pi_1^m, \ldots, \pi_{K^m}^m)$ are estimated by the relative selection-frequencies of the code-vectors in the training stage. Additionally, for each frequency bin, source specific noise variances

Acknowledgement: This work was supported by the Austrian Science Fund (project number P22488-N23).

$\boldsymbol{v}^m = (v_1^m, \ldots, v_D^m)^T$ are obtained by variance estimates of k-means' residual error. The d^{th} frequency bin of the mixture spectrum x_d is assumed to be distributed according to a Gaussian:

$$p(x_d|\mathbf{z}) = \mathcal{N}\left(x_d | w_{z^{a_d},d}^{a_d}, v_d^{a_d}\right), \tag{1}$$

where $\mathbf{z} = (z^1, \ldots, z^M)^T$ is the vector of all codebook indices and $a_d = \arg\max_m (w_{z^m,d}^m)$. Assuming independence among the frequency bins, the likelihood of the mixture spectrum $\mathbf{x}$ is given as

$$p(\mathbf{x}|\mathbf{z}) = \prod_{d=1}^{D} p(x_d|\mathbf{z}). \tag{2}$$

Hence, (2) defines a multi-variate Gaussian distribution, with mean vector $\boldsymbol{\mu} = \max_m (\mathbf{w}_{z^m}^m)$, where the maximum is taken element-wise, and a diagonal covariance matrix $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{v})$, with $\boldsymbol{v} = (v_1^{a_1}, \ldots, v_D^{a_D})$. The assumption that $\boldsymbol{\mu}$ is the element-wise maximum of the sources' code-vectors, follows the MIXMAX approximation [7]. For SCSS, we are interested in the maximum posterior (MAP) solution $\mathbf{z}^*$ of codebook indices:

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} \left[p(\mathbf{x}|\mathbf{z}) \prod_{m=1}^{M} p(z^m) \right]. \tag{3}$$

The MAP solution can be found by exhaustive search, whose computational effort grows exponentially in the number of sources. Alternatively, search heuristics can be applied in order to find a solution with large, but possibly sub-optimal posterior probability [10].

3. FACTORIAL LINEAR-VQ

When we operate in the magnitude spectral domain (i.e. without log), we can approximate the mixture spectrum as the sum over the source spectra: $\mathbf{x} \approx \sum_{m=1}^{M} \mathbf{s}^m$. As in max-VQ, we assume that the d^{th} frequency bin of the mixture is distributed according to a Gaussian, as in (1) and (2). However, applying the sum-approximation, we obtain different mean and variance vectors, namely

$$\boldsymbol{\mu} = \sum_{m=1}^{M} \mathbf{w}_{z^m}^m, \quad \boldsymbol{v} = \sum_{m=1}^{M} \boldsymbol{v}^m, \tag{4}$$

where we used the fact that the sum of Gaussian random variables is again Gaussian. The posterior of this model is given as in (3). The codebooks $\mathbf{W}^m$, the priors $\boldsymbol{\pi}^m$, and the variances $\boldsymbol{v}^m$ are obtained as in section 2, with the only difference, that magnitude spectrograms instead of log-magnitude spectrograms are used for k-means training.

4. TIME-FREQUENCY MASKS

Once we have found the MAP indices $\mathbf{z}^*$, we can interpret the corresponding code-vectors as approximations $\hat{\mathbf{s}}^m$ of the sources' magnitude spectra: $\hat{\mathbf{s}}^m = \exp(\mathbf{w}_{z^m}^m)$, for max-VQ, and $\hat{\mathbf{s}}^m = \mathbf{w}_{z^m}^m$, for linear-VQ. The binary mask (BM) for the m^{th} source is then given as

$$\text{BM}_d^m = \begin{cases} 1, & \text{if } \hat{s}_d^m > \hat{s}_d^l,\ l \neq m, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the BM represents a hard assignment of time-frequency bins to specific sources. To resynthesize time signals, one multiplies the BM with the original complex spectrogram and performs the inverse short-time Fourier transform, i.e. the inverse Fourier transform for each short-time spectrum, followed by an overlap-and-add procedure. Alternatively, we can define a continuous mask (CM), according to a Wiener filter:

$$\text{CM}_d^m = \frac{(\hat{s}_d^m)^2}{\sum_{l=1}^{M} (\hat{s}_d^l)^2}.$$

The CM can take any value between 0 and 1 and represents a soft assignment of frequency bins to sources. Hence, when the goal is to resynthesize the separated signals, we expect a better signal quality using a CM. On the other hand, when the goal is signal analysis or high interference suppression, a BM is preferred. In [11], the estimation of the ideal BM is depicted as the ultimate goal of CASA.

5. EXPERIMENTS

In our experiments we used data from the GRID corpus [9], where we selected speakers 18 and 20 (female), and speakers 1 and 2 (male). For each speaker, we selected 10 random test utterances, while the remaining 490 utterances were used as training data. All speech signals were sampled at 16 kHz, normalized to zero-mean and to unit standard deviation. For the spectrograms, we took frames of 1024 samples with 50 % overlap and applied a hamming window. We trained several codebooks with K codewords, $K \in \{50, 100, 200, 300, 400, 500\}$. We considered all 6 speaker pairs, and for each speaker pair we mixed all 100 possible combinations of test utterances, where we used anechoic, instantaneous mixtures at a mixing level of 0 dB.

5.1. Performance Measures

Vincent et al. [12] proposed four measures in order to evaluate the performance of audio source separation algorithms. These are the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-noise ratio (SNR), and signal-to-artifact ratio (SAR). Let $\hat{s}$ be an estimation of a certain target source s_{target} in time domain, extracted by an algorithm under test. They assumed the following decomposition: $\hat{s} = \hat{s}_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}$, where $\hat{s}_{\text{target}}$ is the desired target signal s_{target} (up to some scale), e_{interf} is the sum of interfering signals (each up to scale), e_{noise} is the known noise signal (up to scale), and $e_{\text{artif}} = \hat{s} - \hat{s}_{\text{target}} - e_{\text{interf}} - e_{\text{noise}}$. In order to estimate these components, we define the matrix $\mathbf{B} = (s_{\text{target}}, s_{\text{int}_1}, \ldots, s_{\text{int}_{M-1}}, s_{\text{n}})$, whose columns contain the target signal, the $M-1$ interfering signals, and the noise signal. Shorter signals are zero-padded in order to have the same length as the longest signal. We calculate the coefficients $(c_{\text{target}}, c_{\text{int}_1}, \ldots, c_{\text{int}_{M-1}}, c_n)^T = \mathbf{B}^\dagger \hat{s}$, where † denotes the pseudo-inverse. The assumed signal components are then given as $\hat{s}_{\text{target}} = s_{\text{target}}\, c_{\text{target}}$, $e_{\text{interf}} = (s_{\text{int}_1}, \ldots, s_{\text{int}_{M-1}})(c_{\text{int}_1}, \ldots, c_{\text{int}_{M-1}})^T$, $e_{\text{noise}} = s_{\text{n}}\, c_n$, and $e_{\text{artif}} = \hat{s} - \hat{s}_{\text{target}} - e_{\text{interf}} - e_{\text{noise}}$. With these signal components at hand, the performance measures are defined as:

$$\text{SDR} := \frac{\|\hat{s}_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}, \tag{5}$$

$$\text{SIR} := \frac{\|\hat{s}_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}, \tag{6}$$

$$\text{SNR} := \frac{\|\hat{s}_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{noise}}\|^2}, \tag{7}$$

$$\text{SAR} := \frac{\|\hat{s}_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2}. \tag{8}$$

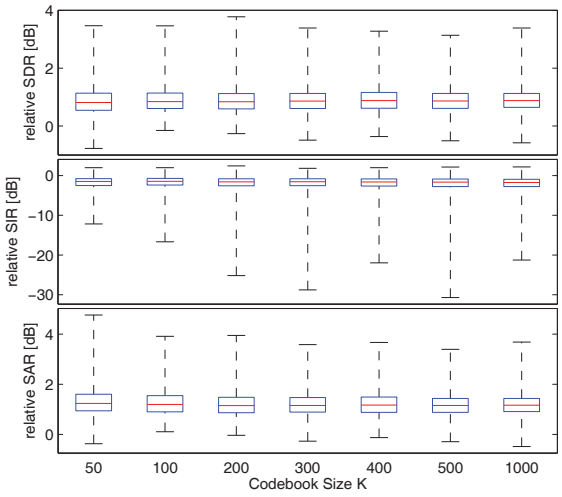

Fig. 1. Relative performance comparison between continuous mask and binary mask as a function of codebook size K. Horizontal red lines correspond to the median, the boxes corresponds to the 25 % and 75 % quantiles, and the whiskers represent the total range.

The SDR corresponds to overall signal quality, since it compares the target signal energy to the energy of all unwanted components. The SIR reflects how well the interfering sources are suppressed in comparison to the target signal. The SNR follows the usual definition, i.e. it compares "useful" signal energy against noise. The SAR measures the amount of artifacts introduced by the separation algorithm.

5.2. Binary versus Continuous Mask

In this section, we study the performance difference between binary and continuous masks. We applied max-VQ and linear-VQ to all our test mixtures, where no noise was added. We summarize our results using *relative* performance between the CM and the BM. The relative SDR is calculated as $\mathrm{SDR}_{\mathrm{rel}} = \frac{\mathrm{SDR}_{\mathrm{CM}}}{\mathrm{SDR}_{\mathrm{BM}}}$, where $\mathrm{SDR}_{\mathrm{CM}}$ and $\mathrm{SDR}_{\mathrm{BM}}$ is the SDR obtained for a CM or a BM, respectively. For the SIR and the SAR, we proceed likewise. We do not consider the SNR here, since no noise was added and the SNR is infinite by definition. We calculated relative performance measures for all mixtures, combining results from max-VQ and linear-VQ, which results in 2400 values for each value of K. Figure 1 shows a box plot of the relative performance. We see that the CM consistently achieves a significantly higher SDR, which corresponds to overall signal quality. The results for the SAR suggest that the higher SDR for the CM stems from a smaller amount of artifacts introduced during resynthesis. On the other hand, the BM suppresses the interfering speaker significantly better than the CM, resulting in a higher SIR. Since a BM represents a more radical decision and a hard assignment of time-frequency bins to sources, it is clear that a BM introduces more artifacts than a smooth CM, while better suppressing the interfering speaker. We note that these results are not very surprising and assume them to be widely known. However, we are not aware of an empirical study explicitly confirming these results.

5.3. Max-VQ versus Linear-VQ

In this section, we compare factorial max-VQ with factorial linear-VQ. For this purpose, we distinguish 4 categories of separated signals: female - female, male - female, female - male, male - male, where the former is the gender of the target speaker, while the latter is the gender of the interfering speaker. Figure 2 compares the median performance of max-VQ and linear-VQ in terms of SDR, SIR and SAR, again for the noise free case. Due to lack of space, we only show the results when a CM is used for resynthesis. The results for the BM are likewise. We see that linear-VQ significantly outperforms max-VQ, especially for small codebooks. For larger codebooks, the performance difference becomes smaller, and in the female-female case for $K = 500$, max-VQ performs slightly better than linear-VQ. Subjective perceptual evaluation by the authors confirmed these results. One reason for the worse performance of max-VQ might be, that logarithmic spectra are spread over $\mathbb{R}^D$, while magnitude spectra fill only the positive orthant. Since k-means aims to minimize the sum of the Euclidean distance errors, the approximation of the log-magnitude spectra is less accurate than when modeled in the linear domain with the same number of cluster-vectors.

Furthermore, we compare max-VQ and linear-VQ in terms of noise-robustness. For this purpose, we contaminated the mixture utterances with additive white Gaussian noise, resulting in SNRs of 20 dB, 10 dB and 5 dB. We used a codebook size of $K = 500$ and a CM for resynthesis. Figure 3 compares the performance of the two systems, dependent on the input noise level. Although the performance of both methods drops significantly with larger noise levels, and both systems fail for SNR = 5 dB, we can state that max-VQ is significantly less robust against noise, especially in terms of SDR.

6. CONCLUSION

We compared the sum approximation for magnitude spectra and the MIXMAX approximation for log-magnitude spectra for VQ-based SCSS. As a technical contribution, we proposed factorial linear-VQ, the linear counterpart of the factorial max-VQ system. The main result of our work is that a seemingly irrelevant choice of features (whether to apply the log or not) has a significant impact on the performance of VQ-based source separation. As a second contribution, we systematically compare the performance of source separation systems using a binary or a continuous time-frequency mask. We can conclude that a binary mask suppresses the interfering sources better than a continuous mask, while the continuous mask introduces less artifacts and achieves a better signal quality. Hence, a binary mask can be advantageous for applications which require an isolated representation of the sources, e.g. automatic speech recognition. However, when sources have to be resynthesized, and high signal quality is required, a continuous mask should be preferred.

7. REFERENCES

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. J. Wiley and Sons Ltd, 2006.

[2] S. Roweis, "One microphone source separation," in *Neural Information Processing Systems*, 2001, pp. 793–799.

[3] ——, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, 2003, pp. 1009–1012.

[4] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[5] S. Rennie, J. Hershey, and P. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Processing Magazine*, vol. 27, pp. 66–80, 2010.

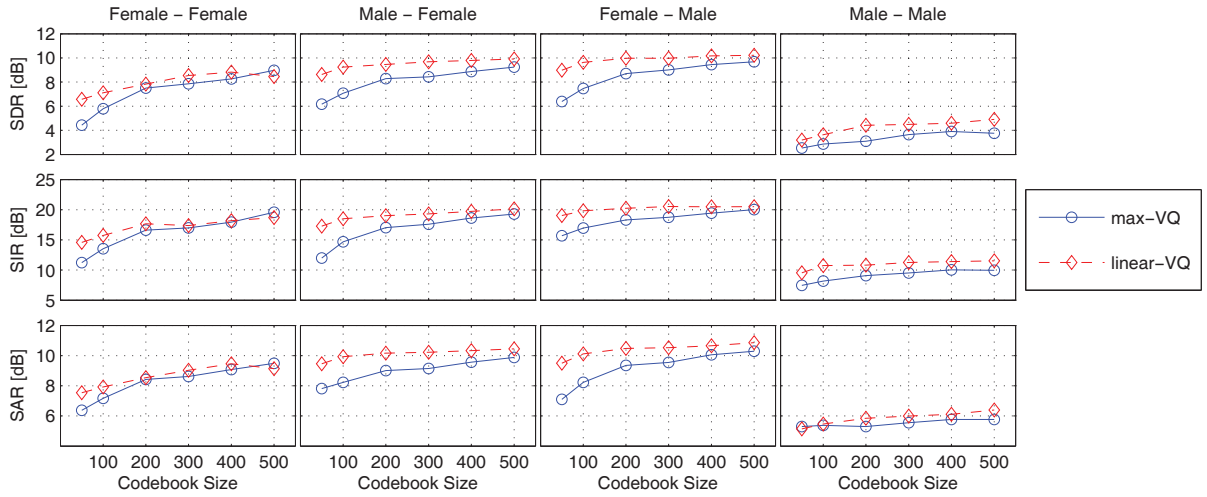

Fig. 2. Median performance of max-VQ and linear-VQ as function of codebook size K. Each column corresponds to a different gender combination of the form: gender of target speaker - gender of interfering speaker. A continuous mask was used for resynthesis.

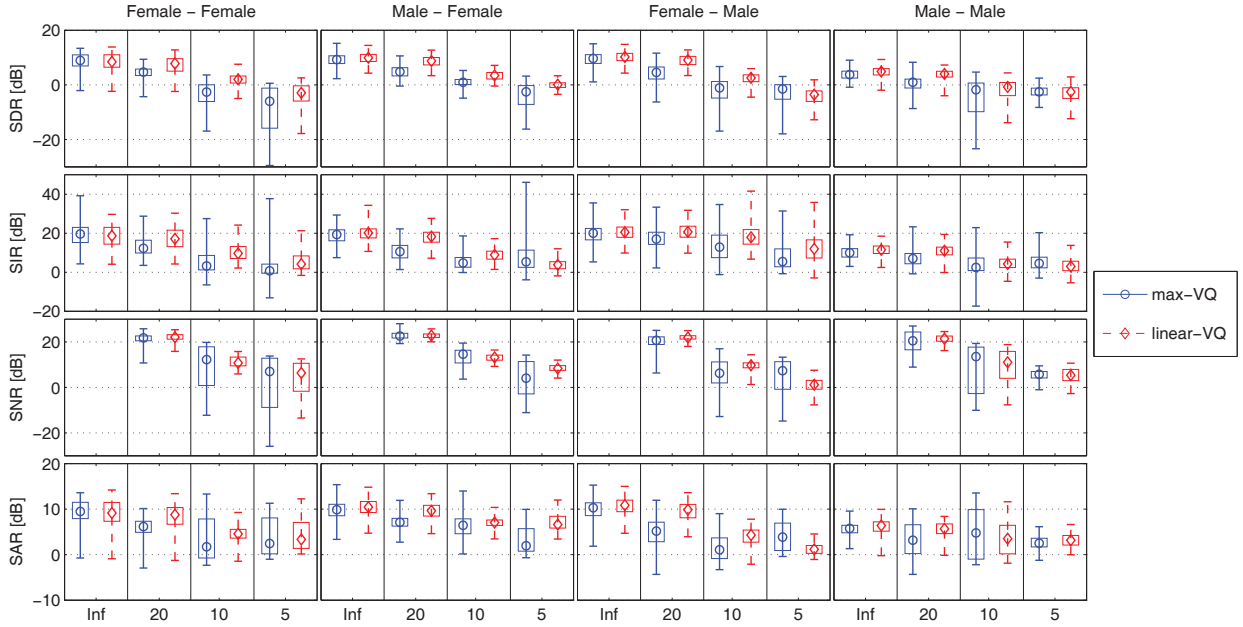

Fig. 3. Performance of max-VQ and linear-VQ for speech mixtures plus additive white Gaussian noise, for codebook size $K = 500$. X-axis: input noise level in SNR [dB]. Markers correspond to medians, boxes to the 25% and 75% quantiles, and whiskers mark the overall range. A continuous mask was used for resynthesis. No SNR is shown for the noise free case, since the output SNR is infinite per definition.

[6] P. Smaragdis, "Approximate nearest-subspace representations for sound mixtures," in *ICASSP*, 2011, pp. 5892 – 5895.

[7] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," in *ICASSP*, vol. 1, 1988, pp. 517 – 520.

[8] M. Radfar, A. Banihashemi, R. Dansereau, and A. Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximization approximation," *Electronic Letters*, vol. 42, pp. 75–76, 2006.

[9] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," in *JASA*, no. 120, 2006, pp. 2421–2424.

[10] M. Stark and F. Pernkopf, "On optimizing the computational complexity for VQ-based single channel source separation," in *ICASSP*, 2010, pp. 237–240.

[11] D. Wang, *Speech Separation by Humans and Machines*. Kluwer Academic, 2005, ch. On ideal binary mask as the computational goal of auditory scene analysis, pp. 181–197.

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