INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Exploiting Conic Affinity Measures to Design Speech Enhancement Systems Operating in Unseen Noise Conditions

Pavlos Papadopoulos, Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, USA

ppapadop@usc.edu, shri@sipi.usc.edu

Abstract

Speech enhancement under unseen noise conditions is a challenging task, but essential for meeting the increasing demand for speech technologies to operate in diverse and dynamic real world environments. A method that has been widely used to enhance speech signals is nonnegative matrix factorization (NMF). In the training phase NMF produces speech and noise dictionaries which are represented as matrices with nonnegative entries. The quality of the enhanced signal depends on the reconstruction ability of the dictionaries. A geometric interpretation of these nonnegative matrices enables us to cast them as convex polyhedral cones in the positive orthant. In this work, we employ conic affinity measures to design systems able to operate in unseen noise conditions, by selecting an appropriate noise dictionary amongst a pool of potential candidates. We show that such a method yields results similar to those that would be produced if the oracle noise dictionary was used.

Index Terms: Non-negative Matrix Factorization, Speech Enhancement, Convex Optimization, Conic Affinity

1. Introduction

The last few years have witnessed an increased demand of speech applications operating in a variety of diverse real life environments including home, vehicles, and outdoor settings. Hence, the ability of speech technologies to operate under different, and often unseen, types of noise is closely related with their overall performance. This has renewed the interest of researchers in speech denoising, and has resulted in the development of methods that are not restricted to specific types of noise. Such schemes include subspace methods with time and spectral constraints [1, 2]. More recently, the community has focused its attention on methods based on Deep Neural Networks (DNN) [3, 4], as well as Nonnegative Matrix Factorization (NMF) [5, 6, 7].

Methods utilizing DNNs for denoising are usually trained by combining clean audio signals with a diverse pool of noises, having the noisy speech signal as the input and its clean version as the target output. These methods are data intensive, since they require a large amount of pairings between clean speech signals and various types of noise at different SNR levels. On the other hand, NMF methods do not suffer from this data dependency; however, they require prior information about the type of noise that corrupts the speech signal. Obviously, this information cannot always be made available, and ways to deal with such issues have been addressed in prior work [8].

In the training phase, NMF-based speech enhancement utilizes magnitude spectrograms to construct spectral representations of the speech and the noise that corrupts the signal, respectively. These spectral representations, also known as dictionaries, are used in the testing phase to enhance the speech signal. This is achieved by expressing the magnitude of the noisy spectrogram (the spectrogram of the signal corrupted by noise) as a conic combination of speech and noise dictionary atoms and subsequently disregarding the part of the noisy spectrogram projected onto the noise dictionary. This is the reason we need to know beforehand the type of noise that corrupts the signal, because without this knowledge we would not be able to construct the noise dictionary that is necessary in the testing phase. However, in certain applications, e.g. enhancing a person's speech while imaged in an MRI scanner, you could design protocols that would allow you capture noise without explicit prior knowledge. The authors in [9] achieved that by capturing MRI noise for a specific amount of time without the human speech interference, and constructed their noise dictionary based on this information. Of course such an approach is not always feasible because many noises in real-life environments do not exhibit the characteristics that are conducive to easy dictionary designs. In our earlier work [8], we proposed two methods–a noise selection scheme and a combined dictionary approach, to overcome these issues. Of particular interest is the combined dictionary approach, where the noise dictionary used in the testing phase is a "concatenation" of the available known noise dictionaries. Although, this approach shows satisfactory results there are cases where it fails. It was noted that a possible solution would be to develop a selection scheme that would choose only "valid" candidates that would be included in the combined dictionary. To achieve that one would need to develop robust mathematical tools to compare those NMF dictionaries.

In this work, we examine how to exploit the geometrical properties of NMF in order to design speech enhancement systems that are able to operate in unseen noise conditions. To that end, we investigate different conic affinity measures [10] which give information regarding how "similar" two convex polyhedral cones are. Given a noisy signal, we use the conic affinity measures to make informed decisions about which noise dictionary to use from a pool of available noises. Once a noise dictionary is selected it is employed in the denoising phase to produce the enhanced signal. Different conic affinity measures have been used to address various challenges, such as image clustering [11], and studying the dynamics of large metabolic networks [12].

We evaluate the performance of our system based on two metrics: Perceptual Evaluation of Speech Quality (PESQ) improvements [13], and segmental-SNR improvements[14]. We show that using such techniques one can design systems that are able to perform in unseen conditions, with comparable performance when a oracle noise dictionary is used, while at the same time avoiding the need of huge amounts of data.

The dictionaries produced by NMF are nonnegative, hence they can be interpreted as generators of convex polyhedral cones in the positive orthant [15]. The dictionary atoms express the extreme rays of the convex polyhedral cone. If the dictionaries

http://dx.doi.org/10.21437/Interspeech.2020-1269

created contain non-extreme rays they can be removed, since the geometry of the cone will remain unchanged. Identifying non extreme rays can be achieved by a simple feasibility test, where we test if the ray can be expressed as a conic combination of the remaining rays. In our experiments we did not encounter such cases. In fact, the enhanced speech spectrogram is a conic combination of the speech dictionary atoms.

The geometrical properties of NMF have been exploited to address various problems in the literature. For example, in [16] an NMF modification based on convexity is proposed and applied in hyperspectral imaging (HSI). The authors in [17] create the dictionary by constructing the conic hull of the training data instead of using an objective function to minimize the reconstruction error [18]. Moreover, given a source, i.e., speech or noise, Kim *et al.* created a set of local dictionaries to capture the source's manifold [19].

The rest of the paper is organized as follows. In Section 2 we give an overview of NMF and provide insights about its geometrical interpretation. In Section 3, we describe the system, as well as the conic affinity measures we employ in our study. In Section 4, we present our experiments, discuss the results, and outline some interesting directions for future research. Finally, in Section 5 we summarize our work and provide our conclusions.

2. Explaining NMF through a geometric approach

Given a non-negative matrix $V \in \mathbb{R}^{K\times N}$, NMF attempts to find non-negative matrices $W \in \mathbb{R}^{K\times L}$ and $H \in \mathbb{R}^{L\times N}$ such that $V \approx WH$[1]. In order to find this approximation, one needs to solve the following optimization problem:

$$\begin{aligned} \underset{W,H}{\text{minimize}} \quad & D(V||WH) \\ \text{subject to} \quad & W \succeq 0,\ H \succeq 0 \end{aligned}$$

where $X \succeq 0$ means that all the elements of X are nonnegative, while $D(\cdot)$ is a separable cost function such that:

$$D(V|WH) = \sum_{k=1}^{K}\sum_{n=1}^{N} d(V_{kn}||[WH]_{kn})$$

where A_{ij}, and $[A]_{ij}$ stand for the element of matrix A at row i and column j. The cost function $D(\cdot)$ that is commonly used is the β-divergence [20].

In the special case of $\beta = 2$, the β-divergence reduces to the Euclidean distance, which is the cost function we use in this work. Since we need to optimize with respect to both W and H, an iterative procedure is used where updates for W and H are alternated until convergence.

In the speech enhancement framework, NMF is applied in the following way. In the training phase, we compute a speech dictionary $W_{speech} \in \mathbb{R}^{K\times L}$, and a noise dictionary $W_{noise} \in \mathbb{R}^{K\times L}$, from their corresponding spectrogram magnitudes, where the design parameters K, and L represent the number of frequency bins and the number of dictionary basis vectors respectively. We assume, without loss of generality, that both the speech and noise dictionaries have the same number of basis vectors L. In the testing phase, we estimate the activation matrix $H_{noisy} \in \mathbb{R}^{2L\times M}$ that best approximates the magnitude spectrogram of the noisy signal $V_{noisy} \in \mathbb{R}^{K\times M}$:

$$V_{noisy} \approx [W_{speech}\ W_{noise}]H_{noisy} \tag{1}$$

where W_{speech} and W_{noise} are fixed and retrieved from the training phase. Finally the enhanced spectrogram magnitude $\hat{V}$ is calculated by:

$$\hat{V} = W_{speech}H' \tag{2}$$

where H' is the $L \times M$ matrix consisting of the first L columns of H_{noisy}, i.e. $H' = [h_1^T;\ h_2^T; \dots h_L^T]$, with h_j^T being the j row of H_{noisy}.

Assuming that the magnitude spectrogram V_{noisy} consists of M frames, then Equations (1) and (2) can be expressed as:

$$v_m \approx [W_{speech}\ W_{noise}]h_m \quad \forall m = 1, 2, \dots, M \tag{3}$$

$$\hat{v}_m = W_{speech}h'_m \quad \forall m = 1, 2, \dots, M \tag{4}$$

where v_m, $\hat{v}_m$ are the m-th frames of V_{noisy} and $\hat{V}$ respectively, and h_m, h'_m the m-th columns of H_{noisy} and H'.

Since the dictionaries W_{speech}, W_{noise} are nonnegative, then by extension their combination $[W_{speech}\ W_{noise}]$, will also contain only nonnegative values. Thus, all these dictionaries can be considered as generators of convex polyhedral cones in the positive orthant [15]. Given a matrix P, a convex polyhedral is the set defined by the conic combination of its columns:

$$\begin{aligned} \Gamma_P &= \left\{ x : x = \sum_j \alpha_j P_j,\ a_j \geq 0\ \ \forall j \right\} \\ &= \{x : x = P\alpha,\ \alpha \succeq 0\} \end{aligned} \tag{5}$$

where P_j are the columns of P, α_j are nonnegative constants, and α a vector whose elements are the α_j values.

Notice that all the elements of h_m in Eq. (3) are nonnegative, since h_m is a column of the nonnegative matrix H_{noisy}. Thus, v_m is the conic combination of the atoms in $[W_{speech}\ W_{noise}]$ according to equation (5). Therefore, in the NMF framework the noisy frame is a point in the cone Γ_C generated by $C = [W_{speech}\ W_{noise}]$.

This insight is crucial for understanding how speech enhancement is achieved in the NMF framework. A noisy frame v_m is expressed as a point in the in the cone Γ_C generated by combining the speech and noise dictionaries. Hence, the noisy frame is described as the conic combination of the extreme rays of Γ_C, or equivalently the conic combination of the atoms from the speech, and the noise dictionary. The result of this process is that the noise dictionary will capture the noise-only information of the frame, separating it from the speech components, by adjusting accordingly the corresponding conic coefficients. Finally, once the noisy frame v_m is decomposed, eq. (3), we retrieve the enhanced frame by keeping only the activations that correspond to the speech dictionary, eq. (4).

The quality of the enhanced signal depends on the ability of the cone Γ_N, generated by W_{noise}, to accurately model the noise components of the signal. Therefore, prior knowledge regarding the type of noise that corrupts the signal is necessary to create an accurate dictionary W_{noise}. However, this is not always feasible, and has attracted research efforts to address this issue. There are various methods proposed in the literature, for

[1] Throughout this work the matrix V, upon which NMF is applied, stands for the magnitude of the spectrogram.

example, [8] uses a noise selection scheme to decide which dictionary to use in the denoising phase, while a similar approach is being followed in [21] for SNR estimation. Other methods attempt to capture the signal information directly without focusing on specific noise types, e.g., such an approach has been used for image clustering in [11].

Based on these principles, we could develop methods that measure "similarity" of convex polyhedral cones, which could guide the design of systems that use some form of noise-selection to compensate for missing noise information.

3. System Description

We will explore four conic affinity measures and their individual performance. The first one exploits the Euclidean distance of a point to a cone, the second one is based on cosine similarity, the third takes into account the truncated Pompeiu-Hausdorff metric, and finally the fourth one uses the ball-truncated volume of the cone.

Consider two cones Γ_A, Γ_B generated by matrices A, and B. We assume without loss of generality that the columns of both matrices act as the extreme rays of the cones they generate. The first affinity measure is defined as the average Euclidean distance of each extreme ray in Γ_A to the cone Γ_B:

$$\delta_d(\Gamma_A, \Gamma_B) \coloneqq \frac{1}{K}\sum_{k=1}^{K} d(a_k, \Gamma_B) \tag{6}$$

where a_k is an extreme ray of Γ_A, K the number of extreme rays and $d(a_k, \Gamma_B)$ the Euclidean distance of a_k to the cone Γ_B. In order to find the required distance we need to solve the following convex quadratic problem:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & ||Bx - a_k||_2^2 \\ \text{subject to} \quad & x \geq 0 \end{aligned}$$

In our case, the cones are generated by the NMF dictionaries. Since the atoms of those dictionaries can have different ℓ_2 norms, we normalize all the atoms to unit ℓ_2 norm in order to have consistent distance values, without affecting the performance in the denoising phase. Notice that smaller values of $\delta_d(\Gamma_A, \Gamma_B)$ indicate that the two cones Γ_A, Γ_B are closer in the multidimensional space they are defined.

The second conic affinity measure is based on pairwise cosine similarity between the extreme rays of cones. For each of the two cones Γ_A, Γ_B, we form random conic combinations of their extreme rays to produce new points within their respective sets.

The result of this "sampling" process are the sets $C_A \subset \Gamma_A$ and $C_B \subset \Gamma_B$. Following this, we find the vectors $a_i \in C_A$ and $b_i \in C_B$ with the maximum cosine similarity:

$$s(a_i, b_i) = \frac{\sum_{m=1}^{M} a_{im} \cdot b_{im}}{\sqrt{\sum_{m=1}^{M} a_{im}^2} \cdot \sqrt{\sum_{m=1}^{M} b_{im}^2}}$$

Subsequently, these vectors are removed from C_A and C_B and we repeat the process. Finally, we compute the average cosine similarity of all pairs:

$$\delta_s(\Gamma_A, \Gamma_B) \coloneqq \frac{1}{|C_A|}\sum_{r=1}^{|C_A|} s(a_r, b_r) \tag{7}$$

where $|C_A| = |C_B|$ is the cardinality of the set C_A, and a_r, b_r are points in the sets C_A and C_B respectively. Notice that $\delta_s(\Gamma_A, \Gamma_B)$ is bounded between 0 and 1 and higher values of $\delta_s(\Gamma_A, \Gamma_B)$ indicate high degree of similarity between the two cones Γ_A, Γ_B.

The Pompeiu-Hausdorff metric is defined as:

$$\delta_{PH} \coloneqq \text{haus}(\Gamma_A \cap \mathbb{B}_n, \Gamma_B \cap \mathbb{B}_n) \tag{8}$$

where $\mathbb{B}_n$ is the closed unit ball in $\mathbb{R}^n$ and

$$\text{haus}(\Gamma_A, \Gamma_B) = \max\left\{\max_{x\in\Gamma_A} d(x, \Gamma_B), \max_{x\in\Gamma_B} d(x, \Gamma_A)\right\}$$

We remind the reader that $d(x, \Gamma_A)$ is the euclidean distance of point x to the cone Γ_A. In order to calculate haus(Γ_A, Γ_B) one needs to solve two conic linear programming problems.

Finally the ball-truncated volume of the cone is defined as:

$$\text{btv}(K) \coloneqq \text{vol}_n(K \cap \mathbb{B}_n), \tag{9}$$

where vol_n stands for the n-dimensional Lebesgue measure.

One way of calculating the expression in (9) is by using the n-dimensional Gaussian measure as shown in [22]. Hence we can write:

$$\frac{\text{btv}(K)}{\text{vol}_n(\mathbb{B}_n)} = \frac{1}{(2\pi)^{n/2}}\int_K e^{-\frac{1}{2}||x||^2} dx \tag{10}$$

We can utilize the above conic affinity measures, along with a diverse pool of available noise dictionaries to design systems operating in unseen noise conditions. For each noise in our pool we calculate the corresponding dictionaries. Subsequently, when the system is presented with a noisy signal, we apply the NMF procedure and produce a "dictionary" representation W_{noisy} and calculate the conic affinity measures of the cone generated by W_{noisy} with those generated from the noise dictionaries.

In this work we test the performance of each affinity measure individually. So the system designed based on the measure defined by (6) will calculate the metric based on W_{noisy} and each of the noise dictionaries and the one that will yield the smallest value will be the selected dictionary. In contrast, the system that is based on (7) will select the noise dictionary that corresponds to the highest value of the metric since higher values of $\delta_s(\cdot\,,\cdot)$ indicate higher degrees of similarity. The system designed based on Pompeiu-Hausdorff metric selects the dictionary corresponding to the lowest value. Finally, the system that uses the ball-truncated volume makes it decision in the following way: First, we form the matrices $[W_{noisy}\ W_m]$, for each of the m noises in our pool. Then we calculate the quantity described in (10) for each of the cones generated from these matrices. Notice that in all cases the dimensionality n is the same thus we can ignore the constant $\text{vol}_n(\mathbb{B}_n)$ when comparing the volumes, while the final value is calculated by "sampling" the cone as we did in the case of cosine similarity.

4. Experiments and Discussion

To perform our experiments we need clean speech utterances to construct speaker-specific dictionaries, as well as a pool of noises with different characteristics. To that end, we use 50 male and 50 female speakers from the TIMIT database [23].

Table 1: *Performance of speech enhancement for five noise types selected from the NOISEX-92 database (white, speech babble, high frequency, machine gun, and factory floor 1). Performance is measured with respect to two metrics: PESQ, and segmental-SNR improvements. In all cases "Oracle" represents the dictionary that corresponds to the noise that corrupts the signals, "Best in pool" corresponds to the dictionary that results in the best performance if we exclude the oracle, and "Combined" is the combined dictionary approach presented in [8]. The other columns correspond to the systems designed based on the conic affinity measures described in (6), (7), (8), and (9).*

	PESQ							**segmental-SNR**						
	δ_d	δ_s	δ_{PH}	btv	Oracle	Best in pool	Combined	δ_d	δ_s	δ_{PH}	btv	Oracle	Best in Pool	Combined
W.	0.586	0.586	0.586	0.586	0.667	0.586	0.712	17.342	17.342	17.342	17.342	19.453	17.342	18.658
S.B.	0.257	0.323	0.298	0.323	0.392	0.323	0.411	5.043	4.642	5.244	4.956	5.437	5.244	3.901
H.F.	0.436	0.458	0.408	0.486	0.511	0.486	0.367	13.181	15.248	11.481	15.101	17.014	15.248	12.023
M.G.	0.482	0.534	0.134	0.534	0.603	0.534	0.417	9.123	10.355	11.541	12.042	13.787	12.042	6.762
F.F.1.	0.422	0.303	0.422	0.378	0.451	0.422	0.311	11.435	12.981	12.981	13.431	15.656	13.431	11.283

Each dictionary is trained using 9 utterances. The test utterances are corrupted with noises from the NOISEX-92 database [24], at SNR levels of 0 dB and 5 dB. The NOISEX-92 database contains 15 types of noise with different characteristics. All the spectrograms were extracted using 25 ms windows with an overlap of 10 ms and 512 frequency bins. Thus, for each speaker and noise type, we have dictionaries of 257 atoms, and all the atoms were normalized to unit length.

Moreover, we examine the ability of the proposed systems to perform in unseen noise conditions with the following experiment. We corrupt speech utterances with a specific type of noise and then we remove it from the noise pool, thus the system cannot pick the type of noise that alters the signal and is forced to pick another noise with "similar" characteristics. For example, if we corrupt an utterance with Military Vehicle noise, this specific noise is removed from the pool, and the system will select one of the remaining 14 types of noise. We enhance the noisy signal with the selected dictionary, and measure the quality of the produced signal in terms of Perceptual Evaluation of Speech Quality (PESQ), and segmental-SNR score improvements. We compare the performance of the proposed systems with the "oracle" dictionary, i.e. the dictionary of the noise that actually corrupts the signal, and the best available dictionary in the noise pool.

The results of this experiment are presented in Table 1. An immediate observation that can be made is that all the systems perform well and often they choose the best available option, and the performance is on par with the oracle noise dictionary, which is the dictionary that corresponds to the type of noise the signal was corrupted with. Obviously, the system is dependent on the size and variability of the noise pool. Notice that for both PESQ and segmental-SNR the systems show satisfactory performance. In addition to that we observe that the combined dictionary does not always gives the best performance, a result that agrees with those in [8]. A combined dictionary could result in a cone that spans a larger area of the orthant, allowing the algorithm to select atoms that would express more accurately the conic combination, however this area could include regions of speech. This could be a possible explanation that justifies the inconsistent results of the combined dictionary approach.

Although we do not present such experiments here due to space limitations, the reader should notice that if we restrict the noise pool to only a few types of noise, or to noises that they all share the same characteristics, it severely limits the ability of the system to select appropriate noise dictionaries to enhance the signal, resulting in poor performance.

Furthermore, these results support our hypothesis that conic affinity measures could be employed to guide the design of NMF-based systems able to operate in unseen noise conditions. This warrants further investigation on utilizing the geometric properties of NMF produced dictionaries to improve the performance of speech technologies. Similar systems found in literature use signal extracted features (e.g., MFCC, filterbanks, etc), [21, 25].

Another interesting observation is that the systems based on different conic affinity measures hold complementary information. This leads to the conclusion that a method combining these conic affinity measures along with others capturing different characteristics of the cones' geometry could yield superior performance.

Our next steps will be focused on two aspects. First, we need to enrich our system with more conic affinity measures that are able to capture information regarding different aspects of the cones than the ones we have already used, for example the orientation of the cone in the positive orthant. Second, we need to investigate methods that that combine the information from different conic affinity measures and take advantage of the complementary information these affinity measures hold. For example, if we consider each affinity measure as an "expert" informing us about its decision, then we could employ schemes from graph theory (e.g., Schulze method) to make more sophisticated and accurate decisions regarding the selected dictionary.

5. Conclusions

In this work we examined four conic affinity measures and how they could be used to design speech processing systems operating in unseen noise conditions. The systems utilize these affinity measures to select a noise dictionary from an available pool of noises. The conic affinity measures attempt to provide a degree of "similarity" between two convex polyhedral cones, enabling us to find an appropriate noise dictionary during the speech enhancement process. The results indicate that these conic affinity measures hold information which allow us to make informed decisions regarding which noise dictionary to use. We measured the performance of our system with respect to two speech enhancement metrics commonly used in the literature, and found that our system selects an appropriate dictionary to enhance the signal. In the future, we plan to incorporate more conic affinity measures that will capture different aspects of the cone geometry, and investigate a more robust noise selection criterion.

6. References

[1] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[2] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[3] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[4] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proceedings of InterSpeech*, 2018.

[5] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4029–4032.

[6] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *INTERSPEECH , Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 411–414.

[7] C. Vaz, V. Ramanarayanan, and S. Narayanan, "A two-step technique for mri audio enhancement using dictionary learning and wavelet packet analysis." in *INTERSPEECH*, 2013, pp. 1312–1315.

[8] P. Papadopoulos, C. Vaz, and S. S. Narayanan, "Noise aware and combined noise models for speech denoising in unknown noise conditions," in *INTERSPEECH*, 2016.

[9] C. Vaz, V. Ramanarayanan, and S. Narayanan, "A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis," in *Interspeech*, 2013, pp. 1312–1315.

[10] P. Papadopoulos, C. Vaz, and S. S. Narayanan, "Exploring the relationship between conic affinity of nmf dictionaries and speech enhancement metrics," in *Proceedings of InterSpeech*, 2018.

[11] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1, June 2003.

[12] D. Grigoriev, S. Samal, S. Vakulenko, and A. Weber, "Algorithms to study large metabolic network dynamics," *Mathematical Modelling of Natural Phenomena*, vol. 10, no. 5, pp. 100–118, 2015.

[13] 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 Proceedings.*, 2001.

[14] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[15] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. MIT Press, 2004, pp. 1141–1148.

[16] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3239–3252, July 2012.

[17] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML)*, 2013.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[19] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 293–297, March 2015.

[20] S. Romain, E. Slim, and R. Gael, "Group nonnegative matrix factorisation with speaker and session variability compensation for speaker identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5470–5474.

[21] P. Papadopoulos, A. Tsiartas, and S. Narayanan, "Long-term snr estimation of speech signals in known and unknown channel conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2495–2506, Dec 2016.

[22] D. Gourion and A. Seeger, "Deterministic and stochastic methods for computing volumetric moduli of convex cones," *Computational and Applied Mathematics*, pp. 215–246, 2010.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," in *Linguistic Data Consortium*, Philadelphia, PA, 1993.

[24] A. Varga and H. J. M. Steeneken, "Assessment for Automatic Speech Recognition II: NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, July 1993.

[25] P. Papadopoulos, R. Travadi, and S. S. Narayanan, "Global SNR estimation of speech signals for unknown noise conditions using noise adapted non-linear regression," in *INTERSPEECH*. ISCA, 2017, pp. 3842–3846.