Interspeech 2018
2-6 September 2018, Hyderabad

Exploring the Relationship between Conic Affinity of NMF Dictionaries and Speech Enhancement Metrics

Pavlos Papadopoulos, Colin Vaz, Shrikanth Narayanan

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

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

Abstract

Nonnegative Matrix Factorization (NMF) has been successfully used in speech enhancement. In the training phase NMF produces speech and noise dictionaries, whose elements are nonnegative, while in the testing phase it estimates a non-negative activation matrix to express the enhanced speech signal as a conic combination of those dictionaries. This nonnegativity property enables us to interpret them as convex polyhedral cones that lie in the positive orthant. Conic affinity could be useful when designing NMF-based systems for unseen noise conditions, which operate by selecting an appropriate noise dictionary amongst a pool of potential candidates. To that end, we examine two conic affinity measures, one based on cosine similarity, while the other is based on euclidean distance from a point to a cone. Moreover, we construct an algorithm to show that conic affinity correlates with speech enhancement performance metrics.

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

1. Introduction

The performance of speech processing applications degrades in the presence of noise. In the last few years, data availability and the increased demand of speech applications operating in real world scenarios resulted in the development of novel denoising methods that are not restricted to specific noise types. 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], as well as Nonnegative Matrix Factorization (NMF) [4, 5, 6].

DNN-based methods utilize large sets of speech and a diverse noise pool to train the network by pairing noisy frames with the corresponding clean ones. On the other hand, NMF methods do not have this extreme data dependency and produce similar results; however, they require prior information about the type of noise that corrupts the speech signal. This type of knowledge cannot always be available, especially if the data are collected from various sources and under varying noise conditions. In [7] the authors propose some methods to alleviate this issue.

Broadly speaking, NMF speech enhancement consists of two phases: training and testing. In the training phase, we use the magnitude spectrograms to construct spectral representations of the speech and the noise that corrupts the signal. This process involves the decomposition of the magnitude spectrogram into two non-negative matrices: a dictionary and an activation matrix. In the testing phase, the speech and noise dictionaries are employed to enhance the noisy signal. This is achieved by expressing the magnitude of the noisy spectrogram as a conic combination of speech and noise dictionary atoms and subsequently disregarding the part of the noisy spectrogram projected onto the noise dictionary.

Since the dictionaries produced by NMF are nonnegative they can be interpreted as convex polyhedral cones in the positive orthant [8], with the dictionary atoms acting as the extreme rays of the cone[1]. In fact, the enhanced speech spectrogram is a conic combination of the speech dictionary atoms.

The geometrical properties of NMF have been exploited to attack various problems in the literature. For example, in [9] an NMF modification based on convexity is proposed and applied in hyperspectral imaging (HSI). The authors in [10] create the dictionary by constructing the conic hull of the training data instead of using an objective function to minimize the reconstruction error [11]. Moreover, given a source, i.e. speech or noise, Kim *et al.* created a set of local dictionaries to capture the source's manifold.

The motivation behind this work is straightforward; given a noise pool, and their NMF cone representations, we investigate conic affinity measures that can be utilized to select an appropriate cone for the denoising phase. Conic affinity measures have been successfully applied in various applications, such as image clustering [12], and studying the dynamics of large metabolic networks [13], to name a few. We demonstrate that there exists a relation between conic affinity and speech enhancement performance in terms of three metrics: Perceptual Evaluation of Speech Quality (PESQ) improvements [14], segmental SNR improvements, and Weighted-Slope Spectral distance (WSS) improvements[15].

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

2. A Geometric Interpretation of NMF

Given a non-negative matrix $V \in \mathbb{R}^{K \times N}$, in our case the magnitude of the spectrogram, the goal of NMF is to find non-negative matrices $W \in \mathbb{R}^{K \times L}$ and $H \in \mathbb{R}^{L \times N}$ such that $V \approx WH$. This approximation is achieved by solving the following optimization problem:

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

[1]If the dictionaries 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.

10.21437/Interspeech.2018-1387

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 $[A]_{ij}$ denote the element of matrix A at row i and column j. A common choice for the cost function is the β-divergence [16], defined as:

$$d_\beta(x||y) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta + \beta x y^{(\beta-1)}) & \beta \in \mathbb{R}\setminus\{0,1\} \\ x\log\frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases}$$

Notice that when $\beta = 2$, the above expression reduces to the Euclidean distance. In traditional approaches, the updates for W and H are alternated until convergence; first, the cost function is optimized for W with H treated as a constant, then the cost fuction is optimized for H with W fixed. In this work, we will use the Euclidean distance ($\beta = 2$) for the cost function. The update equations in this case are:

$$W_{kl} \leftarrow W_{kl}\frac{[VH^T]_{kl}}{[WHH^T]_{kl}}$$

$$H_{ln} \leftarrow H_{ln}\frac{[W^TV]_{ln}}{[W^TWH]_{ln}}$$

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 represents 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. $H' = [h_1^T;\ h_2^T; \dots h_L^T]$, where h_j^T is row j 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'.

By construction, the dictionaries W_{speech}, W_{noise}, and by extension their combination $[W_{speech}\ W_{noise}]$, contain only nonnegative values. Hence, they can be interpreted as generators of convex polyhedral cones in the positive orthant [8]. 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.

Since all the elements of h_m in Eq. (3) are nonnegative (as a column of the nonnegative matrix H_{noisy}), v_m is approximated as a conic combination of the atoms in $[W_{speech}\ W_{noise}]$. Therefore, in NMF the noisy frame is expressed as 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. The noisy frame is decomposed into speech and noise components in the combined speech and noise cone Γ_C. The noise dictionary will capture the noise-only information of the signal, separating it from the speech components. Thus, 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).

It is clear that 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. Hence, it is necessary to have prior knowledge about the type of noise that corrupts the signal. However, this is not always possible, and various methods have been proposed in the literature to address this issue. For example, the authors in [7], use a noise selection scheme to decide which dictionary to use in the denoising phase, while a similar approach has been used for SNR estimation in [17] and image clustering in [12]. Hence, investigating conic affinity measures could guide the design of such systems by selecting the appropriate noise through its cone representation.

3. Conic Affinity Measures

We construct two conic affinity measures: one based on Euclidean distance of a point to a cone, and another on based on cosine similarity.

3.1. Distance of a point to a 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. We define the first affinity measure as the average Euclidean distance of each extreme ray in Γ_A to the cone Γ_B:

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

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. We calculate the distance by solving 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. Normalization of dictionary atoms to unit length does not alter the performance in the denoising phase.

We expect 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.

3.2. Cosine similarity

We calculate the second conic affinity, cosine similarity, through the following procedure. 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. For example, if Γ_A is generated by matrix $A \in \mathbb{R}^{M \times N}$ and every column of A acts as an extreme ray of Γ_A, then for random vectors $z_i = [z_{i1}\ z_{i2} \dots\ z_{iN}]$, where $z_{ij} \geq 0,\ 1 \leq j \leq N$, the point x_i:

$$x_i = z_{i1} \begin{bmatrix} A_{11} \\ A_{12} \\ \vdots \\ A_{M1} \end{bmatrix} + z_{i2} \begin{bmatrix} A_{21} \\ A_{22} \\ \vdots \\ A_{M2} \end{bmatrix} + \dots + z_{iN} \begin{bmatrix} A_{N1} \\ A_{N2} \\ \vdots \\ A_{N2} \end{bmatrix}$$

is part of the cone Γ_A.

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:

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

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.

4. Experiments

In our experiments we use 10 male and 10 female speakers from the TIMIT database [18] to create speaker-specific dictionaries. Each dictionary is trained using 9 utterances. We corrupt utterances of those speakers with noises from the NOISEX-92 database [19] (see Table 1) at an SNR levels of 0 dB and 5 dB. The NOISEX-92 database contains 15 types of noise with different characteristics, such as wideband and narrowband noises as well as stationary and nonstationary noises. Both TIMIT and NOISEX-92 are sampled at 16 kHz.

Table 1: *NOISEX-92 noises*

NOISE TYPES	White (W)
	Pink (P.)
	Speech Babble (S.B.)
	Tank (T.)
	Military Vehicle (M.V.)
	Car Interior (C.I.)
	Destroyer Engine Room (D.E.R)
	Destroyer Operations Room (D.O.P)
	F16 Cockpit (F16)
	Factory Floor 1 (F.F.1)
	Factory Floor 2 (F.F.2)
	High Frequency (H.F.)
	Machine Gun (M.G.)
	Jet Cockpit 1 (J.C.1)
	Jet Cockpit 2 (J.C.2)

The spectrograms that were used to train the dictionaries, for both speakers and noises, were extracted using 25 ms windows with an overlap of 10 ms and 512 frequency bins. For each speaker and noise type, we created dictionaries of 257 atoms, which were normalized to unit length.

In order to demonstrate how the affinity measures presented in Section 3 relate to speech enhancement performance metrics, we perform the following experiment. We corrupt speech utterances with a specific type of noise and enhance the signal through NMF while using different noise dictionaries. For each noise dictionary, we observe its effect on the enhanced signal in terms of Perceptual Evaluation of Speech Quality (PESQ), segmental SNR, and Weighted Spectral Slope (WSS) score improvements. Moreover, for each dictionary we measure the values of the two conic affinity measures with the respect to:

- the "oracle" dictionary; that is, the dictionary that corresponds to the type of noise that corrupts the signal.
- the speaker-specific speech dictionary.

Using this information we formulate a decision rule for selecting a noise dictionary amongst candidates, which takes into account both the relationship between the oracle noise dictionary and the speech dictionary. The dictionary selection rule includes the following steps:

1. Calculate $\delta_s(W_{oracle}, W_i)$, $\delta_d(W_{oracle}, W_i)$, for all noise candidates. W_i is the noise dictionary corresponding to candidate i.
2. Find the set S_s that contains the two noises with highest $\delta_s(W_{oracle}, W_i)$, and the set S_d that contains two noises with the highest $\delta_d(W_{oracle}, W_i)$. If $S_s = S_d$ move to the next step. Otherwise, increase S_s and S_d until they contain at least two common elements.
3. For all elements $n \in S_s \cap S_d$ calculate $\delta_d(W_{speech}, W_n)$ and $\delta_s(W_{speech}, W_n)$.
4. Find element $p \in S_s \cap S_d$ with lowest $\delta_d(W_{speech}, W_p)$ and remove it from $S_s \cap S_d$. Repeat until only one element remains in $S_s \cap S_d$. The dictionary that corresponds to the noise remaining in $S_s \cap S_d$ is the one selected.

Table 2: *Performance of speech enhancement metrics for six noises in the NOISEX-92 database (the abbreviations are expanded in Table 1). For each of the three metrics (PESQ, segmental-SNR, WSS), oracle represents the dictionary that corresponds to the noise that corrupts the signals. Selected in the dictionary selected through the method we presented in Section 4. 2nd Best corresponds to the dictionary that results in the best performance if we exclude the oracle. Finally, Worst corresponds to the dictionary, that results in the lowest performance amongst the 14 noise candidates.*

	PESQ				**segmental-SNR**				**WSS**			
	Oracle	Selected	2nd Best	Worst	Oracle	Selected	2nd Best	Worst	Oracle	Selected	2nd Best	Worst
C.I.	0.5070	0.4086	0.4233	-0.0683	15.1519	11.9478	12.1326	0.4985	31.7743	28.3832	28.3832	-0.3519
D.E.R.	0.4925	0.3574	0.3574	0.0186	4.8504	2.5881	2.7218	0.2596	30.1751	16.8767	16.8767	3.4926
D.O.P.	0.4911	0.4851	0.4851	0.1074	4.7053	4.5913	4.5913	0.1199	19.4574	18.002	18.7383	3.133
F.F.1	0.4072	0.3730	0.39531	0.0634	18.5265	16.7111	18.4446	3.0617	3.7963	3.107	3.107	0.2152
M.G.	0.6643	0.5225	0.5225	0.0452	11.3707	7.9637	7.9851	-0.5457	8.0452	6.9567	7.9817	-0.2944
M.V.	0.5987	0.5577	0.5577	0.0434	18.3689	17.8433	17.8433	1.7797	8.3266	7.9767	7.9767	0.2720
J.C.2	0.7184	0.6505	0.6505	0.0391	13.0912	10.098	13.4052	3.9956	4.9335	3.9286	3.9286	-0.4507

Thus, for each of the 15 noises in NOISEX-92, we use the aforementioned method to select a dictionary from the remaining 14, and compare its performance with the oracle, the second best after the oracle, as well as the dictionary with the worst performance. Notice that the oracle is excluded from the selection process since it would be trivial to identify the oracle dictionary with conic affinity measures. Therefore, the goal of our method would be to select the 2nd best dictionary or one close to its performance. Due to space limitations, we present the results for a subset of noises in Table 2. Specifically, we present results for Car Interior (C.I), Destroyer Engine Room (D.E.R.), Destroyer Operation Room (D.O.P), Factory Floor 1 (F.F.1), Machine Gun (M.G.), and Military Vehicle (M.V.). However, the performance is similar for all the noises in NOISEX-92.

In Table 2, we observe that our method always selects a dictionary whose performance is close to 2nd best, and in some cases equal to 2nd best, for all the speech enhancement metrics. Additionally, notice that the performance of the selected dictionary is never close the worst performing one. These two observations suggest that we can exploit the geometrical properties of the cones to find similar types of noise through dictionary representations instead of using signal extracted features (e.g. MFCC, filterbanks, etc) as in [17], [20].

Moreover, these results indicate that we could use conic affinity measures to design NMF-based systems that will be able to enhance speech signals in unseen noise conditions. For example, using the noisy signal and a pre-trained pool of dictionaries we could exploit conic affinity measures to select the most suitable dictionary to enhance the signal.

However, the conic affinity measures we employed have drawbacks. First, the averaging process involved in the computation of $\delta_s(\Gamma_A, \Gamma_B)$ and $\delta_d(\Gamma_A, \Gamma_B)$ reduces the complex geometry of the cones to a single value. Second, we have not exploited measures that provide information regarding the orientation of the cones in the multidimensional space, for example vectors a, b might have the same cosine similarity with vector c but lie in different places in the positive orthant. The same argument holds true for the Euclidean distance.

Furthermore, the conic affinity measures are calculated separately for the noise and speech dictionaries. A metric that jointly measures the conic affinity between the candidate noise cone and the groundtruth noise and speech cones could be beneficial since it would exploit the geometry of both cones simultaneously.

In conclusion, the results presented in this work warrant further investigation. Interpreting NMF dictionaries as cone generators enables us to exploit their geometrical properties and can potentially lead to improved speech enhancement performance in unseen noise conditions.

5. Conclusions and Future Work

In this work we explored conic affinity measures and their relationship with speech enhancement metrics. We found that using conic affinity measures we can make informed decisions about which dictionary to use in the denoising phase. Using a selection procedure we were able to choose dictionaries that result in overall good performance.

Our next steps will focus on two directions. First, we need to investigate more conic affinity measures that will be able to capture the geometry of the cone in a more detailed manner. Additionally, we need measures that provide information regarding the relative orientation of the cone in the space. Such measures could enable us to make accurate comparisons between two cones.

Finally, we will explore methods to utilize these measures to design speech enhancement systems that will be able to operate in unseen noise conditions. To that end we need to take into account the noisy signal, and explore its relation to the speech and noise candidate dictionaries.

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. 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.

[5] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.