

EXPLORATION OF THE ADDITIVITY APPROXIMATION FOR SPECTRAL MAGNITUDES

Stephen Voran

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado, 80305, USA, svoran@its.bldrdoc.gov

ABSTRACT

The separation of acoustic signals is often accomplished through subtractive decompositions of frequency-domain representations. This is typically enabled by the zero phase approximation or the uncorrelated signals approximation but both of these are very coarse approximations in the mathematical sense. We investigate this disconnect between what works in practice and what is mathematically correct. We conduct a broad search for a domain where the additivity of spectral magnitudes is best satisfied. We apply objective estimators to time-domain reconstructions to characterize the true auditory impact of the magnitude additivity approximation. Our results show the auditory impacts of additivity approximations and allow comparison with the impact of using mixture phase and exact magnitudes in the time-domain reconstruction.

Index Terms— auditory scene analysis, compositional model, noise reduction, noise suppression, source separation, spectral subtraction, speech enhancement

1. BACKGROUND

Separating a mixture of acoustic signals into individual components associated with physical sources is an important problem that has received significant attention over the decades. One may seek to separate a desired signal (often speech) from one or more undesired signals (noise suppression) or to isolate individual contributors (talkers, instruments, environmental sounds) in an acoustic scene for further analysis or transmission (source separation).

The problem is often addressed in the frequency-domain using an additive model for the individual sources. More to the point, subtractive decompositions of frequency-domain representations are effective in practice. In the spectral subtraction algorithm an estimate of the spectral magnitude of the noise is subtracted from an estimate of the observed signal, resulting in an estimate of the spectral magnitude of the desired signal [1]-[3]. A parallel development operates in the magnitude-squared (power) domain [4]. These time-honored practices are effective and have been extensively studied, adapted, and enhanced. A small sampling of this rich body of work can be found in [5]-[11].

More recent insights have led to the constructive composition model for acoustic mixtures [12]-[14]. In this model acoustic atoms and activation functions are combined to additively construct time-frequency representations for acoustic mixtures. Sophisticated iterative algorithms can identify the atoms and activation functions required to subtractively decompose an observed time-frequency pattern in a manner that is largely consistent with the underlying physical sources. The approach is intuitive and effective, and at its core lies the subtraction of spectral magnitudes to separate sources. Indeed, researchers have repeatedly demonstrated that additivity of spectral magnitudes is a pragmatic and effective approximation.

Addition is a very good model for combining multiple acoustic waveforms. But spectral magnitudes add only if they have the same phase values and two squared magnitudes add only if they have phase differences of $\pm \frac{\pi}{2}$. Phase differences tend to be uniformly distributed and approximating all phase deferences as zero or $\pm \frac{\pi}{2}$ is a giant mathematical stretch. This disconnect between what is effective in practice and what is mathematically correct can be unsettling. It is natural to ask how good these approximations are and if there might be a better approximation. A prerequisite question is how to relevantly measure the effect of these approximations.

The question of an application-specific optimal exponent for spectral magnitudes has been studied previously, resulting in multiple answers based on multiple application-specific definitions of optimality (e.g., [6], [9], [10], [15], [16].)

Our approach to these questions is novel. We strip away as many application-specific details as possible in order to focus on the core issue of magnitude additivity. We search for a domain where additivity is best satisfied and we search far beyond the domain of the single fixed exponent. We apply objective estimators to time-domain reconstructions in order to characterize the true auditory impact of the magnitude-additivity approximation. Our results show the auditory impacts of additivity approximations and allow comparison with the impact of using mixture phase and exact magnitudes in the time-domain reconstruction.

We develop a generalized mathematical framework for our work in the next section. Then we present our experiment configuration and results for three specific cases: separation of wide-band speech from a single non-stationary noise signal, separation of two to eight musical instruments, separation of two to eight wide-band speech signals. The paper concludes with discussion of the results.

2. MATHEMATICAL FRAMEWORK

Let the vector \mathbf{x}_i contain the time-domain samples of the i^{th} real-valued source signal, $i = 1$ to N , $2 \leq N$. To obtain a complex-valued spectral representation of a signal we form length M (M even) frames with 50% frame overlap. We then apply the Hann window and the Discrete Fourier Transform (DFT) to each frame. We use the operator F_M to represent these well-known processes:

$$F_M(\mathbf{x}_i) = \{X_{idk}e^{j\phi_{idk}}\}, \quad 1 \leq i \leq N, \quad 1 \leq d \leq D, \quad 1 \leq k \leq K, \quad (1)$$

where i , d , and k are the source, frame, and frequency indices, respectively. Note that F_M^{-1} exists and is comprised of the inverse DFT followed by the overlap-and-add process. Let \mathbf{y} represent the sum of the N sources:

$$\mathbf{y} = \sum_{i=1}^N \mathbf{x}_i, \quad (2)$$

and \mathbf{y} has spectral representation

$$F_M(\mathbf{y}) = \{Y_{dk}e^{j\phi_{y_{dk}}}\}, \quad (3)$$

$$1 \leq d \leq D, \quad 1 \leq k \leq K.$$

(We will drop the frame index d and range for k without ambiguity.) The operator F_M is linear so

$$Y_k e^{j\phi_{y_k}} = \sum_{i=1}^N X_{ik} e^{j\phi_{x_{ik}}}. \quad (4)$$

When signals are added the *complex* spectral representations are added as well.

The use of spectral representations provides an added dimension (frequency) that often greatly facilitates the separation of \mathbf{y} into the constituent sources. For example, it is often possible to estimate a spectral representation for a noise signal and then use this estimate to separate speech from a mixture of speech and noise.

But (4) shows that decomposing a spectral component of \mathbf{y} into N spectral components contributed by the N sources requires decomposing a complex number into N complex numbers. The phase values of audio signals are typically more difficult to estimate than the magnitude values [14]. It is fortuitous that the phase values are also often less critical to auditory fidelity than the magnitude values [17]. As a consequence, it is common (successful complex decomposition efforts [18]-[20] and phase reconstruction efforts [21] notwithstanding) to focus efforts on the magnitude portion of (4) and to (often implicitly) make the zero phase approximation: $\phi_{x_{ik}} = \phi_{x_{jk}}, 1 \leq i, j \leq N$ [1]. With this approximation (4) leads to

$$Y_k^p \approx \sum_{i=1}^N X_{ik}^p, \quad (5)$$

with $p = 1$. But one might equally well choose to argue that the signals $x_i, 1 \leq i \leq N$ are approximately uncorrelated at each frequency k [14]. Under this approximation (5) still applies, but with $p = 2$ [4].

Of course in general, real signals show empirical distributions of $\phi_{x_{ik}} - \phi_{x_{jk}}, i \neq j$ that are nearly uniform. Replacing a uniform distribution with one concentrated at zero (supporting $p = 1$) or $\pm \frac{\pi}{2}$ (supporting $p = 2$) is a very coarse approximation indeed. It follows that (5) is a very coarse approximation for $p = 1$ or 2.

This motivates us to invoke the function G_p to aid in our study of this approximation. G_p is a strictly increasing function that maps non-negative real values to non-negative real values. The function is invertible and is parametrized by the vector \mathbf{p} . G_p is intended to map spectral magnitudes to a domain where additivity is approximately satisfied:

$$G_p(Y_k) \approx \sum_{i=1}^N G_p(X_{ik}). \quad (6)$$

In light of (5), the two special cases defined by $\mathbf{p} = [p]$, $G_p(x) = x^p$, $p = 1$ and 2, are of particular importance in our study.

Let $\{\hat{X}_{ik}\}$ be estimates for $\{X_{ik}\}$, $2 \leq i \leq N$. These estimates can be used with (6) to estimate the spectral magnitudes for the first source:

$$X_{1k} \approx \hat{X}_{1k} = G_p^{-1}([G_p(Y_k) - \sum_{i=2}^N G_p(\hat{X}_{ik})]^+). \quad (7)$$

The function $[\cdot]^+$ retains positive values and replaces negative values with zero to enforce the fact that magnitudes are non-negative.

An estimate of the time-domain signal \mathbf{x}_1 can be constructed using the estimated magnitudes $\{\hat{X}_{1k}\}$ along with the phase values of the mixture signal \mathbf{y} :

$$\mathbf{x}_1 \approx \hat{\mathbf{x}}_1 = F_M^{-1}(\{\hat{X}_{1k}e^{j\phi_{y_k}}\}). \quad (8)$$

In general, this use of “mixture” phase in (8) is a pragmatic solution to obtain phase values for signal reconstruction. But when $N = 2$ and certain additional conditions are met, ϕ_{y_k} is the MMSE estimator of $\phi_{x_{1k}}$ and $e^{j\phi_{y_k}}$ is the MMSE estimator of $e^{j\phi_{x_{1k}}}$ [22].

In the case $N = 2$, we can interpret \mathbf{x}_1 as a desired signal and \mathbf{x}_2 as noise and select $G_p(x) = x$ (zero phase approximation) so that (7) will reduce to the classic spectral subtraction result:

$$\hat{X}_{1k} = [Y_k - \hat{X}_{2k}]^+. \quad (9)$$

We have established that the exact decomposition of a mixture of acoustic signals into individual components requires the decomposition of complex numbers into sums of complex numbers. The pragmatic approach has been to invoke the zero phase approximation or the uncorrelated signals approximation and to then manipulate spectral magnitudes or squared magnitudes accordingly. We have introduced the parametrized function G_p to allow generalization beyond these two options. We will now explore the relative merits of these approaches in terms of the perceived quality of the time-domain reconstructions they produce.

3. SIGNAL PROCESSING EXPERIMENTS

Our experiments use objective estimators to compare the original time-domain signal \mathbf{x}_1 with versions recovered from the mixture \mathbf{y} as we vary G_p . We are seeking a G_p that minimizes the perceptual difference between \mathbf{x}_1 and the recovered version. This approach is akin to minimizing the error in the approximation (6) but it eliminates the problem of finding a relevant measure for that error and it includes the effect of using mixture phase to reconstruct the time-domain signal.

We first define $\hat{\mathbf{x}}_1^P$ which is a reference estimate of \mathbf{x}_1 that characterizes only the effect of using mixture phase in a reconstruction:

$$\hat{\mathbf{x}}_1^P = F_M^{-1}(\{X_{1k}e^{j\phi_{y_k}}\}). \quad (10)$$

$\hat{\mathbf{x}}_1^P$ differs from \mathbf{x}_1 due to phase (P) alone.

In typical real problems \mathbf{y} is observed, Y_k are calculated, and one must somehow estimate X_{ik} in order to recover X_{1k} . Algorithms that estimate X_{ik} form an entire field of study, and their behavior will certainly be influenced by the choice of G_p . We wish to focus on finding a domain where spectral magnitudes are most nearly additive. To do so we must eliminate other sources of variation to the extent possible. Thus we use “perfect estimates” or “oracle values” (i.e., the original known values) for the spectral magnitudes of sources 2 through N to estimate those of the first source:

$$\hat{X}_{1k}^A = G_p^{-1}([G_p(Y_k) - \sum_{i=2}^N G_p(X_{ik})]^+). \quad (11)$$

We then use these estimates to generate an estimate of the time-domain signal \mathbf{x}_1 :

$$\hat{\mathbf{x}}_1^{AP} = F_M^{-1}(\{\hat{X}_{1k}^A e^{j\phi_{y_k}}\}). \quad (12)$$

Note that \hat{x}_1^{AP} differs from x_1 due to additivity (A) and phase (P). The additivity component of the difference can be traced to the fact that (6) is an approximation, not an equality. An ideal outcome would be to find a function G_p that transforms spectral magnitudes to a domain where they are truly additive; (6) would then become an equality, the values of \hat{X}_{1k}^A would exactly match those of X_{1k} , and the signal \hat{x}_1^{AP} would match the signal \hat{x}_1^P . Of course we did not achieve this ideal, but objective estimators did tell us which G_p come closest.

Motivated by (5) our experiments included $G_p(x) = x^p$, $0.5 \leq p \leq 2$. We also considered a family of functions that allow the exponent p to increase or decrease monotonically with log frequency:

$$p = [k, p_0, \alpha, \beta], \quad G_p(x) = x^{p_0 + \alpha(\log_2(k) - \log_2(\beta))}. \quad (13)$$

We included a similar family of functions where a single fixed exponent is used for each octave. In another family of functions we allowed different exponents for different magnitude ranges:

$$p = [\alpha_0, \dots, \alpha_m, p_0, \dots, p_{m-1}], \\ \alpha_i \leq x < \alpha_{i+1} \implies G_p(x) = x^{p_i} - \beta_i, \quad 1 \leq i < m, \quad (14)$$

and the β_i were calculated to enforce continuity at each threshold α_i . We also considered the logistic function:

$$p = [\alpha, \beta], \quad G_p(x) = \frac{1}{1 + e^{\beta - \alpha x}}. \quad (15)$$

We optimized the free parameters in each of these functions for best average results across the breadth of each experiment described below. This optimization was driven by objective estimators and equates to the search for a domain in which spectral magnitudes are maximally additive. In spite of these multi-parameter optimizations, no function consistently outperformed $G_p(x) = x^p$. Thus the only results presented below are for $G_p(x) = x^p$. In Section 4 we describe a specific situation where the function in (13) provides a slight advantage.

3.1. Wideband Speech and Non-Stationary Noise

Our first experiment used $N = 2$ sources: wideband speech and non-stationary noise. We used four noise types: street noise, coffee shop noise, an interfering talker, and the babble of ten talkers. The speech material was recorded by five different female talkers and five different male talkers. Each talker recorded 20 unique Harvard Sentences [23] (English language) for a total of 200 sentences lasting over 10 minutes total.

Perceptually consistent objective estimation of audio quality or distortion has been a subject of significant study over the years [24, 25]. We selected the widely-used Wideband Perceptual Evaluation of Speech Quality (WPESQ) algorithm [26] to estimate the quality of \hat{x}_1^P and \hat{x}_1^{AP} resulting in quality estimates Q_P and Q_{AP} , respectively. This application is outside the stated scope of WPESQ, but our listening checks confirmed the reasonableness of the WPESQ results presented here.

The sample rate was $f_s = 16$ kHz. We experimented with M values (processing frame sizes) ranging from 128 to 1024 and found that results (Q_P and Q_{AP}) showed only very slight sensitivity to this parameter. We then selected $M = 512$ (32 ms).

Fig. 1 shows Q_{AP} for the reconstructed speech for three SNR values, four noise types, and a range of p values. Q_{AP} increases with SNR since the additivity and phase issues disappear as x_2 is

reduced. Q_{AP} shows a minor dependence on noise type and the case of the single interfering talker gives the highest quality. The value of p that maximizes Q_{AP} increases slightly with SNR, but always remains in the neighborhood of 1. After averaging across the four noise types, the maximizing value of p (0.05 resolution) moves from 0.95 at SNR=-10 dB, to 1.15 at SNR=+40 dB.

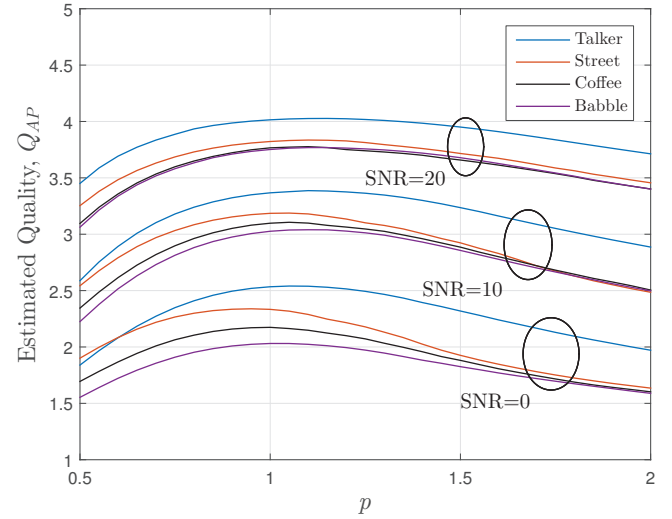


Figure 1: Estimated wideband speech quality as a function of magnitude exponent p for four noise types and three SNRs.

Fig. 2 shows Q_P and three instances of Q_{AP} as a function of SNR. As expected, quality improves with SNR. At every SNR, $p = 1$ gives higher quality than $p = 2$. Optimizing p does not yield a significant increase in Q_{AP} over the case $p = 1$. Comparison of Q_{AP} and Q_P indicates that the major portion of departure from the upper asymptotic quality level (near 4.5) can be attributed to the use of mixture phase. The additional quality drop from Q_P down to Q_{AP} is modest and never exceeds 0.5, which is about 15% of the full quality scale.

3.2. Multiple Musical Instruments

Our musical instrument separation experiment used twelve different multitrack recordings selected from the MTG MASS database [27] and from [28] (converted from $f_s = 44.1$ to 48 kHz). For each recording we created all $N_a! / N!(N_a - N)!$ combinations of tracks, $2 \leq N \leq N_a$ where N_a is the number of tracks (instruments) available, limited to a maximum of eight. The result was between 48 and 98 cases (track combinations) for each value of N . We then reconstructed one of the N sources using oracle values for the spectral magnitudes of the other $N - 1$ sources as in (11) and (12). We repeated this process so that each of the N sources would take the role of x_1 . Our results are based on four minutes of music that cover approximately ten musical genres. We experimented with $M = 1024$, 2048, and 4096, and found that results showed only very slight sensitivity to this parameter. We then selected $M = 1024$ (21.3 ms).

For fullband music the Perceptual Evaluation of Audio Quality (PEAQ) algorithm [29, 30] is somewhat analogous to WPESQ. But PEAQ is intended to quantify distortions much smaller than those found in this work and our listening checks found that PEAQ did not provide useful information in this experiment. PEAQ performs

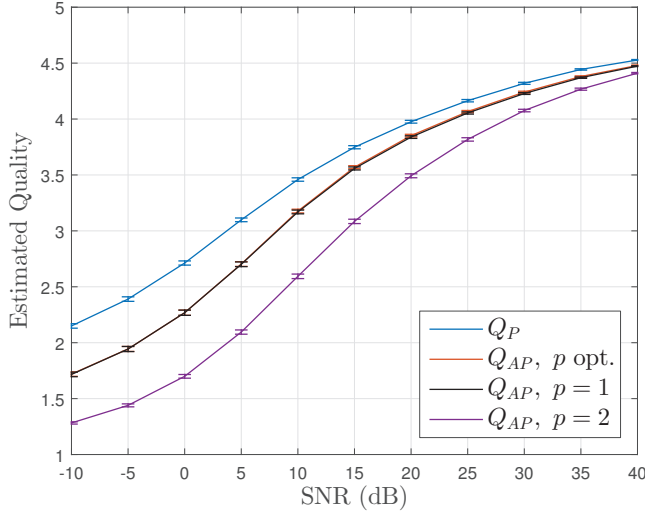


Figure 2: Means and 95% confidence intervals for wideband speech quality vs. SNR.

multiple audio measurements and then combines these through a neural network to create a single output. One of these measurements is average distorted block (ADB). This is a normalized measure of distortion calculated only where that distortion is deemed likely to be detectable by listeners [29, 30]. We have found that ADB has a consistent relationship to perceived distortion in this experiment. Thus we employ ADB and report the perceived distortion in \hat{x}_1^P and \hat{x}_1^{AP} as D_P and D_{AP} , respectively.

Fig. 3 shows D_P and three instances of D_{AP} as a function of the number of sources. Perceived distortion increases modestly as more sources are combined. At every value of N , $p = 1$ gives lower distortion than $p = 2$. Optimizing p to minimize ADB can reduce ADB for small values of N . Comparison with D_P shows that a large portion of the distortion can be attributed to the use of mixture phase. In the case of two sources, D_P is 71% of the D_{AP} result when $p = 1$, and is 84% of the D_{AP} result when p is optimized. When eight sources are combined, D_P shows 63% of the distortion present in the D_{AP} , $p = 1$ case. The optimizing values of p are 1.35, 1.25, and 1.10 for $N = 2, 3$, and 4 respectively. For $N \geq 5$, the optimal value for p is 1.05.

3.3. Multiple Wideband Speech Sources

Our final experiment was similar to the experiment in 3.2 but used the wideband speech of 3.1. We combined $N = 2$ to 8 wideband speech signals (unique talkers) and then reconstructed each one of them using oracle values for the spectral magnitudes of the other $N - 1$ sources. Results were very close to those in 3.1: Q_{AP} values were reliably maximized in the immediate neighborhood of $p = 1$.

4. DISCUSSION AND CONCLUSION

We have searched for a domain where magnitude spectra are maximally additive using a novel and relevant approach that applies objective quality and distortion estimators to time-domain reconstructions. This approach also eliminates the problem of finding a relevant measure for spectral error and it allows for inclusion of the

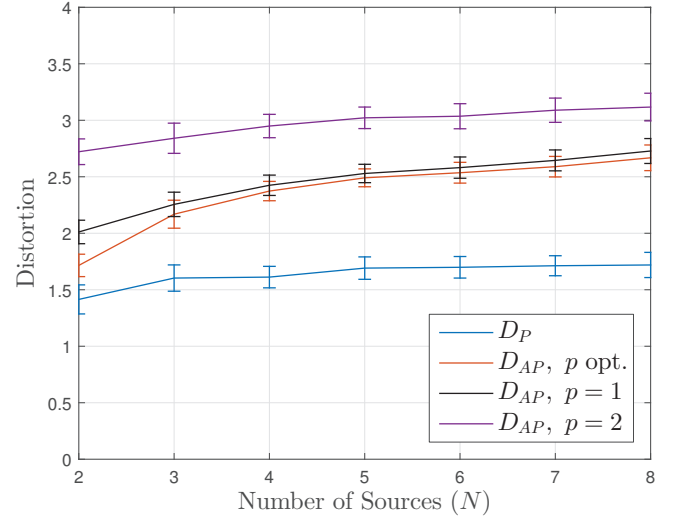


Figure 3: Means and 95% confidence intervals for musical instrument distortion vs. N .

effect of using mixture phase in the reconstruction. We designed our experiments to eliminate as many application-specific factors as possible and they cover many different talkers, sentences, noise conditions, and musical genres. We considered a broad selection of candidate functions for mapping magnitudes to the desired domain.

When the full breadth of this work is considered, the best single domain for adding and subtracting spectral magnitudes is the original domain: $G_p(x) = x^p$ with $p = 1$. This corresponds to approximating phase differences as zero—a great stretch in the mathematical sense, but not far from optimal in the practical sense, at least within the context explored here. Our interpretation is that the variability of magnitude and phase relationships in this class of problems is so great that a single fixed model of these relationships at any level more refined than the highest level (i.e., a single fixed exponent p) is simply not merited and furthermore is not effective.

Our experiments also equip us to judge the magnitude additivity approximation in absolute terms. Comparison of the Q_{AP} and Q_P curves in Fig. 2 or the D_{AP} and D_P curves in Fig. 3 shows that (depending on SNR or N) there is some, but not much, room for improvement in the additivity approximation when mixture phase is used in reconstruction.

Finally, we focus on individual cases. All three experiments showed a weak preference for increasing p values as the separation problem gets easier (higher SNR or fewer sources) and this is consistent with the SNR-driven exponent adaptation formulation given in [10]. When separating $N = 2$ musical sources, $p = 1.35$ has a significant advantage over $p = 1$ (Fig. 3). Separating speech from street noise is another case of interest. Here it is advantageous to use (13) parametrized to increase p from 0.2 at 62.5 Hz to 1.25 at 8 kHz (increasing p by 0.15 each octave). This advantage stems from the fact that street noise is dominated by low-frequency components. That can translate into lower SNR at lower frequencies, thus making smaller p values more suitable at lower frequencies.

We have treated the question of additivity at the most basic level in order to find application-independent results. We expect that these results can provide the basis for further explorations that are individualized for specific applications.

5. REFERENCES

- [1] S. Boll, "Suppression of noise in speech using the SABER method," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Apr. 1978, pp. 606–609.
- [2] —, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, Apr. 1979, pp. 200–203.
- [3] —, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
- [4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, Apr. 1979, pp. 208–211.
- [5] J. Laroche, "Removing preechoes from audio recordings," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1995, pp. 147–150.
- [6] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, Jul. 1998.
- [7] V. Schless and F. Class, "SNR-dependent flooring and noise overestimation for joint application of spectral subtraction and model combination," in *International Conference on Spoken Language Processing*, 1998, pp. 721–725.
- [8] L.P. Yang and Q.J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The Journal of the Acoust. Society of America*, vol. 117, 2005.
- [9] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Noise reduction based on adaptive β -order generalized spectral subtraction for speech enhancement," in *Interspeech 2007*, Aug. 2007, pp. 802–805.
- [10] J. Li, Q.J. Fu, H. Jiang, and M. Akagi, "Psychoacoustically-motivated adaptive β -order generalized spectral subtraction for cochlear implant patients," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2009, pp. 4665–4668.
- [11] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1770–1779, Aug. 2011.
- [12] 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, Mar. 2007.
- [13] P. Smaragdis, "Convolutional speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, Jan. 2007.
- [14] T. Virtanen, J. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, Mar. 2015.
- [15] S. Voran, "Observations on auditory excitation and masking patterns," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1995, pp. 206–209.
- [16] B. King, C. Fevotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *IEEE International Workshop on Machine Learning for Signal Processing*, Sep. 2012, pp. 1–6.
- [17] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, Aug. 1982.
- [18] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2009, pp. 3437–3440.
- [19] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2010, pp. 4206–4209.
- [20] —, "Single-channel source separation using complex matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2591–2597, Nov. 2011.
- [21] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.
- [22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [23] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, Sep. 1969.
- [24] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.
- [25] S. Voran, "Estimation of speech intelligibility and quality," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, M. Vorländer Ed. New York: Springer, 2008, pp. 483–520.
- [26] ITU-T Recommendation P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," Geneva, 2007.
- [27] M. Vinyes, "MTG MASS database," <http://www.mtg.upf.edu/static/mass/resources>, 2008.
- [28] "Nine Inch Nails multitrack sources," <http://ninremixes.com/multitracks.php>, 2005.
- [29] ITU-R Recommendation BS.1387, "Method for Objective Measurements of Perceived Audio Quality," Geneva, 2001.
- [30] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," McGill University, Tech. Rep., Dec. 2003. <http://www.tsp.ece.mcgill.ca>