Speech Enhancement

A Signal Subspace Perspective

Jacob Benesty
Jesper Rindom Jensen
Mads Græsbøll Christensen
Jingdong Chen

Speech Enhancement

A Signal Subspace Perspective

Jacob Benesty
Jesper Rindom Jensen
Mads Græsbøll Christensen
Jingdong Chen

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

AP

Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB
225 Wyman Street, Waltham, MA 02451, USA

First published 2014

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-800139-4

For information on all Academic Press publications
visit our website at **store.elsevier.com**

Transferred to Digital Printing in 2013

Speech Enhancement

CONTENTS

CHAPTER 1

Introduction

In verbal communication, the presence of background noise, such as the sound of a passing car or an air vent, can impact the quality of the speech signal in a detrimental way, something that affects the listener and thus also the communication in several negative ways. Not only may the perceived quality of the speech be harmed, but also its intelligibility may be degraded. Even if only the perceived quality of the speech is affected, this may have a severe impact on the ability of the users to communicate, as exposure to noisy signals may cause listener fatigue. The presence of noise in signals is, though, not only a problem for humans. In speech processing systems, background noise causes additional problems, as such systems often comprise components that are designed under the assumption that only one, clean speech signal is present at any given time. This is, for example, the case for automatic speech recognizers and speech coders. This is typically done to simplify the design of these components, as the underlying statistical models then do not have to account for all possible noise types. Not only does this simplify the training of such models, it also, generally speaking, leads to faster algorithms; but it also renders these components vulnerable to noise.

As we have argued, the presence of background noise is problematic for humans and computers alike, and the problem of dealing with it, which is called speech enhancement or noise reduction, is an important and long-standing problem in signal processing (see, e.g., [1] and [2] for recent surveys), and the search for new and better methods continues today. Speech enhancement algorithms are important components in many systems, where speech plays a part, including telephony, hearing aids, voice over IP, and automatic speech recognizers. Speech enhancement is generally concerned with the problem of enhancing the quality of speech signals. This can, of course, mean many things, but it is often associated with the specific problem of reducing the impact of additive noise, which is also what we are concerned with in the present book. Additive noise occurs naturally in acoustic environments when multiple sources are present, and examples of common noise types are street, car, and babble.

Speech Enhancement. http://dx.doi.org/10.1016/B978-0-12-800139-4.00001-3

Moreover, it can also be caused by intrinsic noise in the sensor system, i.e., from the electrical components. To be more precise, the purpose of speech enhancement is to minimize the impact of the background noise while preserving the speech signal. Hence, there are two performance measures by which the efficiency of speech enhancement methods is compared: speech distortion and noise reduction [3]. These two measures are often conflicting, meaning that if we want to achieve the highest possible noise reduction, then we must accept speech distortion and, similarly, that if we cannot accept any speech distortion, then our ability to perform noise reduction will be hampered. An extreme example of this is the maximum signal-to-noise ratio filter [1] which achieves the highest possible noise reduction but at the cost of severe speech distortion.

The history of noise reduction can be traced back to the work of Wiener [4], i.e., to the very early days of signal processing. Due to the importance of the problem in particular in speech applications, many different solutions have been proposed over the years, and much time and effort is still devoted to the problem today. The problem is often broken into two sub-problems, namely the problem of finding a function to be applied to the observed signal so as to extract the desired signal, i.e., the speech signal, and the problem of finding the information that this function depends on. If we restrict ourselves to linear filters, then the first sub-problem is the problem of finding the optimal filter, i.e., a filter design problem. If the criterion for optimality is the mean-square error, then the so-called Wiener filter is the solution. This filter requires knowledge of the noise statistics (or the speech statistics), and the second sub-problem is then that of finding those statistics, often in the form of the noise correlation matrix or its power spectral density. In the past decade, most work seems to have focused on the second sub-problem, e.g., [5–9], under difficult conditions when the noise is nonstationary. This book is, however, concerned with the first sub-problem, which is determining the function that should be applied to the observed signal. This problem has, though, also seen some important new contributions regarding optimal filtering in the past few years, including [3, 10, 11].

In the literature, one can find many (seemingly) different attempts at solving the problem of speech enhancement, and at the time a new method is published, it is often not clear how exactly it relates to other, existing methods, often because it is either not clear exactly what problems are being solved, or that the problems are stated in different ways whose relation is difficult to ascertain. In fact, it appears that the retrospective process of relating methods may take decades, if it ever occurs. When listing existing classes of methods for speech enhancement, spectral subtraction, (optimal) linear filtering, statistical model-based approaches, and subspace methods are typically mentioned. Indeed, these are also the names of the chapters in the book [2]. The focus in the present book is on the class of methods generally known as optimal filtering, of which the classical Wiener filter is a special case. However, in this book, we will show how speech enhancement using the principles of subspace-based methods can be cast as an optimal filtering problem. As such, the present book unifies what has previously been considered two competing principles of speech enhancement in one framework. As a consequence, it is both possible to combine the benefits of the subspace methods and optimal filtering methods and to analyze and compare the performance of the various approaches analytically.

1.1 HISTORY AND APPLICATIONS OF SUBSPACE METHODS

The development of the subspace-based methods for speech enhancement took a quite different route than the more traditional speech enhancement methods based on the theory of stochastic processes (e.g., linear filtering methods), and it can, therefore, be quite difficult to understand similarities and differences between the methodologies. In that connection, the curious reader might wonder what exactly the distinguishing characteristics of subspace-based enhancement methods are. Subspace-based methods are a class of methods that take their starting point in linear algebra, i.e., they are based on the notions of subspaces and the properties of vectors and matrices. Simply put, they are based on the idea of decomposing the correlation matrix of the observed signal using an eigenvalue-type decomposition and then, from this, find a basis for the part of the space that contains the desired signal (called the signal subspace) and a basis for the part that contains only noise (called the noise subspace).

Subspace methods have a rich history in signal processing, not only for speech enhancement. In fact, much of the early work focused on problems such as parameter estimation, model order estimation, low-rank approximations, etc. Perhaps the earliest example of a subspace method for parameter estimation is Pisarenko's method [12] for sinusoidal parameter estimation. Later followed more, and probably the most famous, subspace methods for the same problem (although cast as the equivalent problem of determining spatial frequencies in arrays) such as the MUltiple SIgnal Classification (MUSIC) method [13, 14] (see also the later papers [15, 16]), of which Pisarenko's method is a special case, and the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) method [17].

Since then, several variations, improvements, and generalizations have followed, including root MUSIC [18], modified MUSIC [19], min-norm [20], unitary ESPRIT [21], and (weighted) subspace fitting [22] (on this matter, see also the tutorial [23]). MUSIC exploits the orthogonality of the signal and noise subspaces while ESPRIT is based on exploiting the structure of the involved matrices, more specifically, their shift-invariance. In [24], it was shown how the model order can be determined statistically from the ratio between the arithmetic and geometric means of the eigenvalues in combination with model selection criteria (this was essentially based on the same derivations as [15]). Later, the ideas behind subspace methods lead to the more general ideas of reduced-rank signal processing [25] and low-rank adaptive filters [26]. More recently, it has been shown that the model order estimation problem can be solved not only based on the eigenvalues (as in [25]) but also by exploiting subspace orthogonality [27] and shift-invariance [28].

The roots of subspace-based enhancement methods can be traced back to [29], although that work appears at first glance to also deal with parameter estimation, focusing on frequency estimation using linear prediction. However, in the paper, enhanced signals are reconstructed via the singular value decomposition of the data matrix, and, hence, the first subspace-based signal enhancement method was born. Much of the early work focused on the simple white noise case [30–32] and later the more general case of colored noise was treated in detail in [33, 34] and later in [35]. In much of this work, the subspace-based enhancement problem is seen as a reduced-rank matrix approximation problem, wherein matrix decompositions are used to obtain a low-rank approximation of a data matrix. Since this approach has its roots in numerical linear algebra, the problem is then often seen as a deterministic one, where the realizations are being approximated, unlike statistical approaches and linear filtering based on stochastic processes. Although connections between matrix decompositions and estimation theory do exist [37] and a filtering interpretation of subspace methods was given in [36], the stated problems are quite different in nature.

At this point, it should be noted that within the general class of subspace methods, various decompositions have been used, like the eigenvalue decomposition, the singular value decomposition, and the Karhunen-Loève transform (e.g., [38, 39]), and some operate on correlation matrices, others on Toeplitz or Hankel data matrices. These are, however, mathematically equivalent, but their numerical properties and memory requirements may differ. Interestingly, triangular decompositions have also been considered more recently [40]. For more on the actual implementation of the various matrix decompositions and their properties, we refer the interested reader to [41]. As many real-time applications require not only fast computations but also sample-by-sample updates, fast methods for computing a basis for the signal or noise subspaces with time-recursive updates, so-called subspace trackers, have been developed [42–46]. Considering that fast and time-recursive implementations of linear filtering approaches are readily available, this is quite important in making the subspace-based methods practical.

1.2 SPEECH ENHANCEMENT FROM A SIGNAL SUBSPACE PERSPECTIVE

We will now go into a bit more detail about how subspace methods work. The simplest possible incarnation of a subspace method for enhancement is perhaps that of a projection of the observed signal onto a subspace known to contain the desired signal (and noise), i.e., the signal subspace. Then, any noise that may lie elsewhere, i.e., in the noise subspace, is removed and noise reduction is achieved without harming the speech signal. A number of questions then arise. First, how do we know that the desired signal, i.e., a speech signal, occupies only a subspace of the full space? Second, and if so, how do we identify this subspace?

To answer the first question, one can look to some commonly used models of speech signals. One such model is the harmonic model, which has a long and rich history in speech processing, specifically for modeling voiced speech. In that model, the speech signal is modeled as a sum of harmonically related sinusoids. For a specific number of harmonics, say C, such a model is well known to have a correlation matrix of rank $P = 2C$ (for the real case). Hence, for any $M \times M$ correlation matrix with $M > P$, the harmonics will lie in a subspace of dimension P. For signals that occupy the full space (whose correlation matrix have rank M), the justification for subspace methods can be found using the theory of low-rank approximation [25]. This theory states that the best rank r approximation of a $M \times M$ matrix (with $r < M$) is obtained by using the r largest singular values and the corresponding singular vectors, and the error (measured using the Frobenius norm) incurred by this is given simply by the remaining, small singular values. This applies, for example, when the number of harmonics of voiced speech exceeds the chosen dimension of the correlation matrix, and for autoregressive processes, which are often used as a model of unvoiced speech.

Returning now to the second question, i.e., how to identify the signal and noise subspaces, there are several ways in which this can be done. When the noise is white, the problem is particularly simple. In that case, the two subspaces can be identified from the eigenvalue decomposition of the observed signal correlation matrix, by simply sorting the eigenvectors according the magnitude of their eigenvalues. Then, the eigenvectors corresponding to the r largest eigenvalues span the same subspace as the desired signal, assuming that its correlation matrix is also rank r. Not only that, they form an orthonormal basis for that space (as the correlation matrices are symmetric by definition), and the eigenvectors corresponding to the remaining eigenvalues form an orthonormal basis for the noise subspace. It then also follows that the two sets of eigenvectors are orthogonal to each other. When the noise is colored, a pre-whitening step has to be included in the process, either as explicit pre-processing, e.g., in the form of filtering, or as part of the eigenvalue decomposition (see, e.g., [40]). More specifically, the appropriate decomposition is that of the generalized eigenvalue decomposition. Seen in a more general way, subspace-based speech enhancement can be seen as a modification of the eigenvalues via a diagonal so-called gain matrix. This way, subspace-based enhancement works by first transforming the signal vector, then applying the gain matrix, after which the signal vector is transformed back.

1.3 SCOPE AND ORGANIZATION OF THE WORK

The purpose of the present book is to unify the approaches of subspace-based enhancement and linear filtering, two approaches that have previously been considered separate classes of methods, and study the combined approach, both analytically as well as experimentally. We start out the book by introducing the general concept with diagonalization of the speech correlation matrix with the eigenvalue decomposition in Chapter 2. We introduce the basic signal model, along with all assumptions, and the basic problem formulation, and define some important quantities and the most important performance measures, namely input and output signal-to-noise ratios (SNRs), the noise reduction factor, the speech reduction factor, and speech distortion index (which are measures of speech distortion). We then proceed to derive several optimal rectangular filtering matrices based on the eigenvalue decomposition, namely the maximum SNR, Wiener, minimum variance distortionless response (MVDR), tradeoff, and linearly constrained minimum variance (LCMV) filters, something that we will continue to do for the various cases considered in the book. In Chapter 3, we then extend these principles to joint diagonalization of the speech and noise correlation matrices, i.e., using generalized eigenvalue decompositions, and analyze their performance. The problem of single-channel speech enhancement in the time domain is then addressed in Chapter 4 using the proposed framework. This is done in two cases: first for a rank-deficient speech correlation matrix, a case that applies, as previously explained, for voiced speech, and then for a full-rank speech correlation matrix. It is then demonstrated how to extend the principles from single-channel to multichannel speech enhancement in Chapter 5, still in the time domain. The generalization to multiple channels turns out to be somewhat complicated, but it leads to an approach that takes spatial information into account. In Chapter 6, the same problem is addressed, but this time in the frequency domain. This leads to a particularly simple solution for the binaural noise reduction problem. Chapter 7 explores a different problem yet, namely that of determining the speech (or signal) subspace using a promising, Bayesian approach based on the Stiefel manifold and the Bingham distribution. Finally, we study the performance of the various filters in simulations in Chapter 8. This is done using synthetic speech signals modeled using a set of harmonically related sinusoids and an autoregressive process, representing voiced and unvoiced speech, respectively. The former model results in speech correlation matrices that are rank deficient while the latter results in full-rank correlation matrices.

REFERENCES

[1] J. Benesty, J. Chen, Y. Huang, I. Cohen, Noise Reduction in Speech Processing, Springer-Verlag, Berlin, Germany, 2009.

[2] P. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Boca Raton, FL, 2007.

[3] J. Benesty, J. Chen, Optimal Time-domain Noise Reduction Filters–A Theoretical Study, Springer Briefs in Electrical and Computer Engineering, Springer-Verlag, 2011.

[4] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications, M.I.T. Press, Boston, MA, 1949.

[5] S. Rangachari, P. Loizou, A noise estimation algorithm for highly nonstationary environments, Speech Commun. 28 (2006) 220–231.

[6] I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging, IEEE Trans. Speech Audio Process. 11 (2003) 466–475.

[7] T. Gerkmann, R.C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay, IEEE Trans. Speech Audio Process. 20 (2012) 1383–1393.

[8] R.C. Hendriks, R. Heusdens, J. Jensen, U. Kjems, Low complexity DFT-domain noise PSD tracking using high-resolution periodograms, EURASIP J. Adv. Signal Process. 2009 (2009) 15.

[9] J.R. Jensen, J. Benesty, M.G. Christensen, S.H. Jensen, Enhancement of single-channel periodic signals in the time-domain, IEEE Trans. Audio Speech Lang. Process. 20 (2012) 1948–1963.

[10] M.G. Christensen, A. Jakobsson, Optimal filter designs for separating and enhancing periodic signals, IEEE Trans. Signal Process. 58 (2010) 5969–5983.

[11] J.R. Jensen, J. Benesty, M.G. Christensen, S.H. Jensen, Non-causal time-domain filters for single-channel noise reduction, IEEE Trans. Audio Speech Lang. Process. 20 (2012) 1526–1541.

[12] V.F. Pisarenko, The retrieval of harmonics from a covariance function, Geophys. J. Roy. Astron. Soc. 33 (1973) 347–366.

[13] R.O. Schmidt, Multiple emitter location and signal parameter estimation, in: Proceedings of the RADC, Spectral Estimation Workshop, 1979, pp. 243–258.

[14] G. Bienvenu, Influence of the spatial coherence of the background noise on high resolution passive methods, Proc. IEEE ICASSP 4 (1979) 306–309.

[15] G. Bienvenu, L. Kopp, Optimality of high resolution array processing using the eigensystem approach, IEEE Trans. Acoust. Speech Signal Process. 31 (1983) 1235–1248.

[16] R.O. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. Antennas Propag. 34 (1986) 276–280.

[17] R. Roy, T. Kailath, ESPRIT – estimation of signal parameters via rotational invariance techniques, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 984–995.

[18] A. Barabell, Improving the resolution performance of eigenstructure-based direction-finding algorithms, Proc. IEEE ICASSP 8 (1983) 336–339.

[19] P. Stoica, K.C. sharman, Maximum likelihood methods for direction-of-arrival estimation, IEEE Trans. Acoust. Speech Signal Process. 38 (1990) 1132–1143.

[20] R. Kumaresan, D.W. Tufts, Estimating the angles of arrival of multiple plane waves, IEEE Trans. Aerosp. Electron. Syst. 19 (1983) 134–139.

[21] M. Haardt, J.A. Nossek, Unitary ESPRIT: how to obtain increased estimation accuracy with a reduced computational burden, IEEE Trans. Signal Process. 43 (1995) 1232–1242.

[22] M. Viberg, B. Ottersten, T. Kailath, Detection and estimation in sensor arrays using weighted subspace fitting, IEEE Trans. Signal Process. 39 (1991) 2436–2449.

[23] H. Krim, M. Viberg, Two decades of array signal processing research: the parametric approach, IEEE Signal Process. Mag. 13 (1996) 67–94.

[24] M. Wax, T. Kailath, Detection of the number of signals by information theoretic criterion, IEEE Trans. Acoust. Speech Signal Process. 33 (1985) 387–392.

[25] L.L. Scharf, The SVD and reduced rank signal processing, Signal Process. 25 (1991) 113–133.

[26] P. Strobach, Low-rank adaptive filters, IEEE Trans. Signal Process. 44 (1996) 2932–2947.

[27] M.G. Christensen, A. Jakobsson, S.H. Jensen, Sinusoidal order estimation using angles between subspaces, EURASIP J. Adv. Signal Process. 2009 (2009) 11.

[28] R. Badeau, B. David, G. Richard, A new perturbation analysis for signal enumeration in rotational invariance techniques, IEEE Trans. Signal Process. 54 (2006) 450–458.

[29] D.W. Tufts, R. Kumaresan, Singular value decomposition and improved frequency estimation using linear prediction, IEEE Trans. Acoust. Speech Signal Process. 30 (1982) 671–675.

[30] J.A. Cadzow, Signal enhancement-a composite property mapping algorithm, IEEE Trans. Acoust. Speech Signal Process. 30 (1988) 671–675.

[31] M. Dendrinos, S. Bakamidis, G. Carayannis, Speech enhancement from noise: a regenerative approach, Speech Commun. 10 (1991) 45–57.

[32] B. De Moor, The singular value decomposition and long and short spaces of noisy matrices, IEEE Trans. Signal Process. 41 (1993) 2826–2838.

[33] Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement, IEEE Trans. Speech Audio Process. 3 (1995) 251–266.

[34] S.H. Jensen, P.C. Hansen, S.D. Hansen, J.A. Sørensen, Reduction of broad-band noise in speech by truncated QSVD, IEEE Trans. Speech Audio Process. 3 (1995) 439–448.

[35] P.S.K. Hansen, Signal subspace methods for speech enhancement, Ph.D. dissertation, Techn. Univ. Denmark, Lyngby, Denmark, 1997.

[36] P. Hansen, S. Jensen, FIR filter representations of reduced-rank noise reduction, IEEE Trans. Signal Process. 46 (1998) 1737–1741.

[37] T.W. Anderson, Asymptotic theory for principal component analysis, Ann. Math. Stat. 34 (1963) 122–148.

[38] U. Mittal, N. Phamdo, Signal/noise KLT based approach for enhancing speech degraded by colored noise, IEEE Trans. Speech Audio Process. 8 (2000) 159–167.

[39] A. Rezayee, S. Gazor, An adaptive KLT approach for speech enhancement, IEEE Trans. Speech Audio Process. 9 (2001) 87–95.

[40] P.C. Hansen, S.H. Jensen, Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: survey and analysis, EURASIP J. Adv. Signal Process. 2007 (2007) 24.

[41] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[42] B. Yang, Projection approximation subspace tracking, IEEE Trans. Signal Process. 41 (1995) 95–107.

[43] D. Rabideau, Fast, rank adaptive subspace tracking and applications, IEEE Trans. Signal Process. 44 (1996) 2229–2244.

[44] R. Badeau, G. Richard, B. David, Sliding window adaptive SVD algorithms, IEEE Trans. Signal Process. 52 (2004) 1–10.

[45] R. Badeau, B. David, G. Richard, Fast approximated power iteration subspace tracking, IEEE Trans. Signal Process. 53 (2005) 2931–2941.

[46] X. Doukopoulos, G. Moustakides, Fast and stable subspace tracking, IEEE Trans. Signal Process. 56 (2008) 1452–1465.