Sparse Time-Frequency Representation of Speech by the Vandermonde Transform

Christian Fischer Pedersen and Tom Bäckström

International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Germany
{christian.fischer, tom.backstrom}@audiolabs-erlangen.de

Abstract

Efficient speech signal representations are prerequisite for efficient speech processing algorithms. The Vandermonde transform is a recently introduced time-frequency transform which provides a sparse and uncorrelated speech signal representation. In contrast, the Fourier transform only decorrelates the signal approximately. To achieve complete decorrelation, the Vandermonde transform is signal adaptive like the Karhunen-Loève transform. Unlike the Karhunen-Loève, however, the Vandermonde transform is a time-frequency transform where the transform domain components correspond to frequency components of the analysis window.

In this paper we analyze the performance of sparse speech signal representation by the Vandermonde transform. This is done by applying matching pursuit and comparing with sparse representations based on dictionaries with Fourier, cosine, Gabor and Karhunen-Loève atoms. Our results show that Karhunen-Loève yields the best sparse signal recovery; however, this is not strictly a time-frequency transform. Of the true time-frequency transforms, Vandermonde is the most efficient for sparse speech signal representation.

Index Terms: Vandermonde transform, sparse speech representation, orthogonal matching pursuit

1. Introduction

Efficient speech processing algorithms require efficient speech signal representations. Efficiency of a representation often include the ability to represent a given signal in a sparse and decorrelated manner in such a way that the underlying signal physics are not only preserved, but also tractable.

A sparse signal representation consists of only a small number of signal describing components which yields a compact representation that allows for fast processing. In such a representation the components are often orthogonal; hereby, each component is able to describe as much as possible of the signal’s variation. Orthogonal representations are also called decorrelated as component decorrelation follows from orthogonality. Decorrelation is not only attractive to sparsity, but also processing. As the components are orthogonal, they may be processed individually without effecting the others. Another desired representation property is that the signal physics, e.g. frequency components, are recognizable and tractable to enable physically motivated processing. Hereby, any existing information about the signal’s physics and origin may serve as basis for the processing.

Within speech processing, many algorithms rely on a sparse signal representations. For example, speech enhancement by sub-space analysis relies on the hypothesis that when applying the Karhunen-Loève transform a sparse signal representation is obtained [1]. Similarly, in speech coding it is assumed that the residual of the linear predictive filter is sparse, whereby it can be encoded accurately by a sparse algebraic codebook [2]. It is therefore fair to expect that improvements in sparse speech representations will be useful in many areas of speech processing.

The objective is thus to design a speech signal representation, which is sparse, decorrelated, and physically motivated. One approach to find a suitable sparse representation is to search the power set of basis vectors for the smallest subset of vectors that is able to yield an accurate representation. However, a brute force search like that is NP complete. Therefore, various adaptive approximation techniques to obtain a sparse representation in an efficient manner have been proposed, e.g. basis pursuit (BP) [3], matching pursuit (MP) [4], and orthogonal matching pursuit (OMP) [5]. All of these methods utilize the notion of cleverly selecting a set of basis vectors out of a large dictionary.

Another approach is to design a signal adaptive set of basis vectors, which are orthogonal and devised such that every successive vector explains as much as possible of the signal’s variation. The Karhunen-Loève transform provides, based on eigenvalue decomposition of the autocorrelation matrix, such a basis. By ordering the eigenvalues decreasingly, it can be shown that every subsequent eigenvector is orthogonal to the previous and that each eigenvector is designed to explain as much as possible of the signal’s variation [6]. Thus, this approach is, with some limitations [7], optimal.

The Karhunen-Loève transform does not, however, provide a physically motivated representation domain. The eigenvectors do not in general describe any physical property or quantity of the signal. The discrete Fourier transform, on the other hand, does provide a physically motivated signal description. In fact, signal estimations by the Karhunen-Loève and Fourier transforms converge when the signal length tend to infinity [8]. Unfortunately, for finite-length signals, and especially short signals, the Fourier transform does not fully decorrelate the signal. This impedes the performance of speech processing algorithms where signals can only be assumed stationary for about 20 ms at a time.

We have recently presented a novel time-frequency transform, the Vandermonde transform, which provides both decorrelation of the input signal as well as a physically motivated representation in the form of frequency components [9]. It corresponds to a warped discrete Fourier transform; warped in the sense that the frequency components are not necessarily uniformly distributed.

Since the Vandermonde transform by design decorrelates the signal and provides a physically interpretable representation, the purpose of the present study is to investigate sparsity. More precisely, the purpose is to analyze the performance of sparse speech representation by the Vandermonde transform.
This is done by applying orthogonal matching pursuit and comparing with sparse representations based on dictionaries with Fourier, Cosine, Gabor and Karhunen-Loève atoms.

The remainder of this paper is organized as follows. Section 2 briefly outlines MP and OMP and section 3 presents the recently proposed Vandermonde transform. Section 4 describes the dictionaries used in OMP for the estimation performance comparison. In section 5 the conducted numerical experiment is described and the results are presented in section 6. In the closing section, section 7, the results are discussed along with future perspectives.

2. Orthogonal Matching Pursuit

In this section, matching pursuit and orthogonal matching pursuit are outlined.

2.1. Matching Pursuit

Basic matching pursuit is an iterative and greedy method to decompose a finite signal in a Hilbert space, $x \in \mathcal{H}$, into a linear combination of signal correlate basis vectors known as atoms [4]. It is greedy in the sense that it determines a locally optimal solution per iteration in an effort to yield a globally optimal solution in the end; a global optimum is not guaranteed by this strategy. The atoms, $g$, have unit length and are selected from a finite and over complete set called a dictionary $D = \{g_\gamma\}_{\gamma \in \Gamma} \subseteq \mathcal{H}$ where $\gamma \in \Gamma$ denotes the parameters that define the atom. Dictionaries are usually over complete as they contain more atoms than necessary to yield an acceptable estimation. The estimation can be formulated as:

$$x(n) = \sum_{m=0}^{M-1} \alpha_m g_{\gamma_m}(n) + R^M(n) = x^M(n) + R^M(n)$$

(1)

Iterating this procedure yields the $M$-term approximation

$$x^M(n) = x(n) - R^M(n) = \sum_{m=0}^{M-1} (R^m(n), g_{\gamma_m}) g_{\gamma_m}$$

Normally, the iterations continue until a stopping criterion is met, e.g. a fixed ceiling on the number of iterations or an acceptable level of energy in the residual.

2.2. Orthogonal Matching Pursuit

In addition to MP, OMP performs an orthogonalization of the constructed basis, $\{g_{\gamma_m}, \alpha_m\}$, at the end of each iteration [11]. This is done to overcome the inherent sub-optimality of an iteration-wise truncated, i.e. finite $M$, MP at the cost of added computations. Let $V_M = \text{span}\{g_{\gamma_m}\}_{m=0}^{M-1}$, then the $M$-term approximation $x^M(n)$ is least-squares optimal iff $R^M(n) \in V_M^\perp$, i.e. iff the $M$-term residual lies in the span of the orthogonal complement of $V_M$ [11]. Basic MP only guarantees that the residual is orthogonal to the last atom chosen, $R^M(n) \perp g_{\gamma_M}$, thus, MP approximations are in general sub-optimal which decreases the rate of estimation convergence. In the OMP orthogonalization at the end of each MP iteration all the coefficients, $\alpha$, of the hitherto constructed basis are recomputed via a least-squares minimization

$$\min_{\{\alpha_m\}_{m=0}^{M-1}} \|x(n) - \sum_{m=0}^{M-1} \alpha_m g_{\gamma_m}\|^2$$

The minimization ensures that $R^M(n) \in V_M^\perp$ and that $x^M(n)$ is least-squares optimal given $\{g_{\gamma_m}, \alpha_m\}_{m=0}^{M-1}$. Hence, the coefficients are recomputed to find the orthogonal projection of $x(n)$ onto $\{g_{\gamma_m}\}_{m=0}^{M-1}$. This approach is similar to the sparse minimization problem

$$\min ||\alpha||_0 \text{ subject to } ||x - G\alpha^T||^2$$

Although OMP improves MP, OMP still selects atoms based on the MP criteria explained in section 2.1, and since MP does not pose an orthogonality constraint on the dictionary, the atoms chosen by the greedy MP approach - and thereby by OMP - may be correlated. Hence, MP does not only select atoms that lie in $\text{span}(V_M)$. This has a negative impact on the rate of convergence of OMP.

In this study we compare estimation performances based on different dictionaries. Some of the dictionaries are orthogonal per design, i.e. Karhunen-Loève and Vandermonde, whereas others are not, i.e. Fourier, cosine, and Gabor. Thus, the results of this study emphasize the importance of applying transforms that yield an orthogonal signal decomposition.

3. Vandermonde Transform

We have recently presented the Vandermonde transform, which simultaneously provides decorrelation, like the Karhunen-Loève transform, and a time-frequency representation of the input signal, like the Fourier transform [9]. Specifically, decorrelation of the input signal corresponds to diagonalization of its autocorrelation matrix. The autocorrelation matrix of a signal $x$ is defined as the expectation of correlation, $R = E[x x^H]$, and it is a Hermitian positive definite Toeplitz matrix. The autocorrelation matrix can be decomposed by various methods, e.g. [6]
By applying the transform \( y = V^{-H}x \) on the signal, where \((V^{-H})\) denotes the inverse Hermitian, we obtain a representation \( y \), which is uncorrelated since its autocorrelation matrix is diagonal

\[
E[yy^H] = E[V^{-H}xx^HV^{-1}] = V^{-H}RV^{-1} = \Sigma
\]

The proposed transform is based on Vandermonde factorization of Toeplitz matrices; the matrix \( V \) has Vandermonde structure

\[
V(v) = \begin{bmatrix}
v_0^1 & \cdots & v_0^{N-1} \\
v_1^1 & \cdots & v_1^{N-1} \\
\vdots & \ddots & \vdots \\
v_{N-1}^1 & \cdots & v_{N-1}^{N-1}
\end{bmatrix}
\]

where the bases of the exponential series have unit length \( ||v_k|| = 1 \). It follows that \( V \) corresponds to a warped discrete Fourier transform, i.e. it is a time-frequency transform where the analysis frequencies are not necessarily uniformly distributed on the unit circle.

4. Design of Dictionaries

We want to evaluate signal decomposition by the novel Vandermonde transform basis; therefore, it is compared against a selection of well-known transform bases. This section describes the bases and how MP dictionaries are designed.

4.1. Vandermonde Transform Basis

The complex-valued Vandermonde transform basis function \( v_{\Omega \varphi}(n) = e^{-i\Omega \varphi} \) has the real-valued counterpart \( \text{Re}(e^{-i\Omega \varphi}) = \cos(-\omega_k n + \varphi) \). Allowing for phase shifts we get the real-valued basis function

\[
v_{\Omega \varphi}(n) = \cos(-\Omega n + \varphi)
\]

where \(-\pi \leq \Omega \leq \pi \) and \( \varphi \in \{0, \frac{\pi}{2}\} \). At \( \varphi = \frac{\pi}{2} \) the imaginary part of the complex-valued basis function is represented. The frequency is sampled in accordance with the Vandermonde factorization, i.e. at the \( N \) frequencies that coincide with the line spectral frequencies pertaining to the palindromic polynomial of a line spectrum pair decomposition. This leads to a dictionary \( D_{VT} \) with \( N \) atoms.

4.2. Karhunen-Loève Transform Basis

The Karhunen-Loève transform (KLT) basis for a finite signal \( x \) is defined as the eigenvectors of the autocorrelation matrix \( R = E[xx^H] \). The main challenges of a KLT basis are the computational time complexity, \( O(n^3) \), of diagonalizing \( R \) and how to interpret the signal’s physical characteristics in the transform domain. On the other hand, a KLT basis is the minimum entropy basis for \( x \) among all other orthonormal bases \([12]\), i.e. it attains the minimum mean squared reconstruction error of \( x \). In resemblance with the Vandermonde transform basis, the KLT basis is computed by diagonalization:

\[
R = QAQ^{-1}
\]

where \( Q \) is a \( N \times N \) matrix with eigenvector columns and \( \Lambda \) is a diagonal matrix with corresponding eigenvalues. Hence, the dictionary \( D_{KLT} \) has \( N \) atoms.

4.3. Fourier Transform Basis

The bilateral discrete-time Fourier transform of \( x = (x_n)_{n \in \mathbb{Z}} \) is defined as \( X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \) where \( 0 \leq \omega \leq 2\pi \). The complex-valued basis function \( f_{\omega}(n) = e^{-j\omega n} \) has the real-valued counterpart \( \text{Re}(e^{-j\omega n}) = \cos(-\omega n) \), i.e. allowing for phase-shifts we get the real-valued basis function

\[
f_{\omega \varphi}(n) = \cos(-\omega_k n + \varphi)
\]

For a signal sequence of length \( N \) we get \( N \) frequency sampling points equidistributed on the unit circle in the complex plane by \( \omega_k = \frac{2\pi k}{N}, \ 0 \leq k \leq N - 1 \). The phase parameter \( \varphi \) is in \( \{0, \frac{\pi}{2}\} \). Again, at \( \varphi = \frac{\pi}{2} \) the imaginary part of the complex valued basis function is represented. This leads to a dictionary \( D_{DFT} \) with \( N \) atoms.

4.4. Cosine Transform II Basis

The Discrete Cosine Transform II basis can be stated as \([13]\):

\[
c_{\omega}(n) = \sqrt{\frac{2}{N}} \cos\left(\frac{\omega (n + \frac{1}{2})}{N}\right)
\]

where \( \omega_k = \frac{2\pi k}{N} \) for \( k = 1, 2, \ldots, N - 1 \). If \( k = 0 \), then \( c_{\omega}(n) = \frac{1}{\sqrt{N}} \). As the frequency sampling is equidistributed at \( N \) points, a dictionary \( D_{DCT} \) with \( N \) atoms is created.

4.5. Gabor Transform Basis

The Gabor Transform is a short-time Fourier Transform with a modulated Gaussian window function. The real-valued Gabor basis function can be stated as \([14, 15]\):

\[
g_{p}(n) = \frac{K_{p}}{\sqrt{\Omega}} \exp\left(-\frac{\pi(n - u)^2}{\sigma^2}\right) \cos(2\pi\omega(n - u) + \varphi)
\]

where the quadruple \( p = (s, u, \omega, \varphi) \) is Gabor function parameters which determine how a basis vector is scaled and positioned in time, frequency and phase respectively. \( K_p \) is a normalization factor such that \( ||g_p||^2 = 1 \).

The parameter settings used in the present study are: \( s = u = 2^q \) where \( 1 \leq p \leq a, \omega = \frac{1}{2}^q \) where \( 1 \leq f \leq a \) and \( q = 2.6 \); thereby, the range of \( \omega \) is \([0; \frac{1}{2}] \); \( q \) is chosen heuristically. The phase parameter varies within \( 0 \leq \varphi \leq \frac{\pi}{2} \) or \( \frac{\pi}{2} \) equidistributed points. To balance the dictionary size with the others: \( a = \sqrt{20N} \). This leads to a dictionary \( D_{Gabor} \) with \( 10N \) atoms; hence, this dictionary contains more atoms than the others. This has been necessary to get the Gabor signal estimation to converge fast enough to be comparable.

5. Numerical experiment setup

The objective with this experiment is to compare the performance of sparse speech signal representation via Vandermonde, Karhunen-Loève, Fourier, Cosine, and Gabor basis functions respectively. OMP is applied in the comparison and the basis functions constitute dictionary atoms, i.e. one dictionary per transform. All dictionaries, except Gabor, are of same size and are constructed to yield a fair and valid comparison. The OMP have only been allowed to run for 25 iterations; thus, all the resulting estimations are based on 25 atoms from the respective dictionaries.

The data material for the experiment is a mono recording of native German clean speech sampled at 12.8 kHz. Rectangular and non-overlapping windowing of length \( N = 256 \) is em-
ployed. A total of 400 consecutive windows constitute the data material. OMP is carried on each of the windows separately.

6. Results

The results in figure 1 illustrate the estimation accuracies in terms of average retained signal energy as a function of the number of included basis vectors, i.e. atoms, from the dictionaries. The average is found across all windows in the data set. The retained energy of signal $x$ is the proportion of retained energy for each OMP iteration, i.e. $\frac{||\alpha_k||_2^2}{||x||_2^2}$ where $\alpha_k$ is the expansion coefficients after the $k$-th iteration.

The results in figure 2 illustrate the estimation residuals’ energy for a worst-case range of windows. For each of the estimations, the windows are sorted in energy descending order. The log residual energy is computed as $\log_{10}(\sum_w (x_w - \hat{x}_w)^2 / \sum_w x_w^2)$ where $w$ is a window index and (’) denotes estimation.

Figure 3 illustrates signal-to-estimation ratios for a representative range of analysis windows. The ratios’ mean and standard deviation for all of the windows in the data set are:

$$
\begin{align*}
(\mu, \sigma)_{KLT} &= (0.35, 0.47) \text{ dB} \\
(\mu, \sigma)_{VT} &= (0.44, 0.60) \text{ dB} \\
(\mu, \sigma)_{DCT} &= (0.51, 0.65) \text{ dB} \\
(\mu, \sigma)_{DFT} &= (0.54, 0.65) \text{ dB} \\
(\mu, \sigma)_{GT} &= (0.89, 1.09) \text{ dB}
\end{align*}
$$

7. Discussion

Karhunen-Loève and Vandermonde shares the benefit of being signal-adaptive which allows the transforms to yield decompositions that are closely related to the underlying signal. This is in contrast to all of the other transforms in this study. Our results show that Karhunen-Loève yields the best sparse speech signal representation; however, this is not strictly a time-frequency transform. Of the true time-frequency transforms in the comparison, Vandermonde is the most efficient for sparse speech signal representation.

Also, it is interesting to compare the Vandermonde estimation with the Fourier estimation as the transforms are very closely related:

$$
\lim_{||D_{DFT}||_0 \to \infty} D_{VT} \subset D_{DFT}
$$

For a future study it would be interesting to investigate the perceived speech quality improvement as a function of basis vectors used in the Vandermonde signal decomposition.

In conclusion, the present study demonstrates that of the true time-frequency transforms in the comparison, Vandermonde is the most efficient for sparse time-frequency representation of clean speech signals.
8. References


