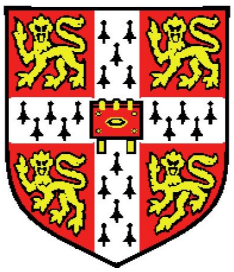
DRAFT COPY ONLY

The detection, classification and restoration of impulses in real-time audio



Jens Enzo Nyby Christensen

Darwin College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

Degree date yet to be decided

Contents

DRAFT COPY ONLY

Chapter 1

Literature review

The literature review presented in this chapter is split into three major sections. Classification, Detection and Restoration. These sections contain the literature review that formed that basis of the different chapters of this document. The classification section will cover literature reviewed for the work undertaken in chapters **??** and **??** while the detection and restoration sections will cover work undertaken in chapters **??** and **??**.

1.1 Classification

1.1.1 Computational auditory scene analysis (CASA)

The computational auditory scene analysis (CASA) approach to source separation attempts to mimic the auditory scene analyzing (ASA) capabilities of the brain by computational means. [140] state: "One may define the CASA problem as the challenge of constructing a machine learning system that achieves human performance in ASA". In an attempt to make CASA systems more *biologically relevant* many CASA systems limit their scope to either monaural or binaural input signals and hence replicating the ASA problem where only the signal from the two ears are present.

More specifically CASA solutions aim to identify certain auditory object, such as characterizing note objects in terms of harmonicity, correlated modulation, duration of sinusoidal partials and common onset, and arranging these into streams using psycho-acoustic cues[25]. Proximity of time-frequency objects within the audio mixture can cause problems for CASA whereas ASA is able to retain a reasonable

degree of separation of sources. These shortcomings of CASA have thus been the focus of much research. Examples of this is the integration of evidence derived from multiple grouping principles at several levels of abstraction[38], matching of timbre features learned on solo excerpts [57],[65],[24] (based on a Gaussian mixture model (GMM) classifier). More recently the overlap probability of frequency components has been introduced to improve algorithm performance[102].

CASA methods still suffer from a few drawbacks though. The brain's own ASA system is able to localize sources in a natural acoustic environment through what is known as the *precedence effect*. In environments where a sound is reflected multiple times before reaching the listener, the brain is able to give *precedence* to early arrival of sounds, such as the direct sound, and hence evaluate the true location of the sound. This precedence rule between grouping cues can be hard to assess [140], which is also why current methods are restricted to non-reverberant mixtures [136].

In the recent paper, [107], the authors present a CASA system for separating speech signals in the presence of other speech or noise. Their approach proceeds by tracking time-frequency units across time frames in 2 different stages. Firstly they use harmonicity to segregate the voiced portions of individual sources for each time frame using multipitch tracking. Secondly the time-frequency units are grouped across the time frames by utilizing knowledge from the speaker characteristics. The system proposed in [107] mainly focusses on features such as onset/offset and periodicity which are not specific necessarily specific to the target source or even speech and as such this system works well for a variety of speech separation cases with a variety of noise scenarios.

1.1.2 Independent Component Analysis (ICA)

One of the most widely known and used methods for blind source separation (BSS) is independent component analysis (ICA). ICA proposes to separate a multivariate signals into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals. For M observations and N sources it is a requirement for ICA to work that $M \geq N$. In general ICA can be defined as the linear transformation $\mathbf{A}$ of observed data $\mathbf{x} = [x_1, \ldots, x_M]$

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1.1}$$

where $\mathbf{s} = [s_1, \ldots, s_N]$ is the separated sources or components [137].

Traditionally $\mathbf{A}$ would then be referred to as the mixing matrix. The source signals $\mathbf{s}$ can feasibly be recovered using $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$ with $\mathbf{A}$ the invertible matrix. $\mathbf{A}^{-1}$ is then called the un-mixing matrix [110].

A very popular implementation of the ICA method is the FastICA algorithm as proposed in [47]. Often mixture models are modeled as *convolutive* and hence solving for the mixing matrix $\mathbf{A}$ in the time-domain will require labor-intensive convolution while some methods propose solutions to the inherent permutation problem for the frequency domain equivalent of the FastICA [86]. It has been reported that this approach has also produced good separation results.

A more recent paper [22] presents two spatio-temporal extensions to the FastICA method and reports significant improvements over previous methods when applied to multichannel recordings of two- and three-source speech mixtures in a room environment. It was also noted that this proposed method was simple, fast and did not require parameter tuning in order to obtain good separation performance.

1.1.2.1 Convolutive mixtures

Due to extensive filtering imposed upon real world signals by their environment and propagation delays, instantaneous mixtures are rarely encountered. Most real world mixtures tend to be of convolved mixtures [110].

Consider again the model from equation (1.1) but this time consider a mixture $\mathbf{x}$ recorded in a real environment where the mixture can be approximated by a convolutive mixture of the source signals in the time domain,

$$\mathbf{x}(t) = \mathbf{A}(t) * \mathbf{s}(t). \tag{1.2}$$

Applying the Fourier transform (FT) yields:

$$\hat{\mathbf{x}}(\omega) = \hat{\mathbf{A}}(\omega)\hat{\mathbf{s}}(\omega), \tag{1.3}$$

where $\hat{\mathbf{x}}(\omega), \hat{\mathbf{A}}(\omega)$ and $\hat{\mathbf{s}}(\omega)$ are the FT of $\mathbf{x}(\omega), \mathbf{A}(\omega)$ and $\mathbf{s}(\omega)$ respectively [48].

It is noted that if **A** in equation (1.1) is seen as a matrix of FIR filters instead of scalars, the resulting multiplication will be equivalent to convolution and the notation will be consistent. This notation is called FIR algebra notation [67].

This time-frequency approach to ICA has been proposed by a variety of authors [72],[48],[110] and specifically [48] has been successful in dealing with the inherent arbitrary scaling and permutation problems of the ICA approach applied in a frequency domain context [152].

The recent paper [152] attempts to avoid the whitening of outputs which plagues conventional implementations of ICA for convolutive mixtures by enforcing pairwise independence rather than mutual independence. This approach is based on the findings that for blind separation the two are generally equivalent.

1.1.3 Independent Subspace Analysis (ISA)

Independent subspace analysis (ISA) is based on ICA but relaxes the constraints $M \geq N$ and the statistical independence of source signals. In practice ISA is a generalization of ICA in the sense that instead of requiring statistical independence between components ISA requires statistical independence between groups of components. ISA can also been described as multidimensional independent component analysis and hence the independent components are required to be of the same dimension k. In the case where $k = 1$, ISA is equivalent to ICA [119]. Despite this relaxing of restrictions of the ICA approach implementations of ISA has still been restricted by fixed group sizes or semi-parametric models. An even more general approach has been proposed which introduces the concept of irreducible independent components and give an identifiability result for this general, parameter-free model together with an arbitrary subspace size algorithm [119].

A previous study managed to implement ISA models for audio source separation and introduced the *ixegram* as a measure space for grouping independent basis components [13]. The separation and grouping results from this experiment suggest that the technique can perform separation of source signals without parametric model fitting or prior knowledge of the composition of input data.

1.1.4 Non-Negative Matrix Factorisation (NMF)

Since the concept of non-negative matrix factorization (NMF) was introduced independently by [97] and [71] the method has gained some popularity within the field of polyphonic transcription and the source separation field [112],[111]. NMF was originally proposed as an alternative to k-means clustering and principal component analysis (PCA) for data analysis and compression. The original formulation of NMF starts with a non-negative $M \times N$ matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ where the goal is to approximate $\mathbf{V}$ as a product of two non-negative matrices so that

$$\mathbf{V} \approx \mathbf{WH} \tag{1.4}$$

where $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$, $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ and $R \leq M$ such that the error of reconstruction is minimized [111].

Good results have generally been reported for extraction of percussive sounds in mono recordings, but for low intensity sounds the algorithms generally perform worse and suffer from distortion and lost information inherent in the mixing and analysis process [111]. Other results have shown NMF to consistently obtain better separation results than ISA and ICA [137].

1.1.5 Hierarchical Bayesian Models

The task in Bayesian source separation is to infer N source signals $s_{k,n} \equiv \mathbf{s}$ given M observed signals $x_{k,m} \equiv \mathbf{x}$, where $n = [1, \ldots, N]$ and $m = [1, \ldots, M]$. Here $k = [1, \ldots, K]$ is an index which may correspond to time or to expansion coefficient of the sources in a linear transform domain. A Hierarchical Bayesian Model for the inference of the source signals can hence be formulated as

$$p(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_x} \int p(\mathbf{x}|\mathbf{s}, \Theta_m) p(\mathbf{s}|\Theta_p) p(\Theta_p) p(\Theta_m) d\Theta_m d\Theta_p. \tag{1.5}$$

Here the conditional distribution is characterized by the conditional distribution $p(\mathbf{x}|\mathbf{s}, \Theta_m)$, where Θ_m denotes the collection of mixing parameters such as the mixing matrix, observation noise and variance, etc. and the prior term $p(\mathbf{s}|\Theta_p)$ describes the sources via their own prior parameters Θ_p [14].

It is worth noting that there is a considerable computational cost associated with a full Bayesian inference, via MCMC or variational Bayes (VB) for Bayesian source separation. Good results were reported in a study attempting Bayesian source separation using an MCMC approach to inference [32]. Also, the authors of this study point out the computational burden of the MCMC approach compared to the alternative expectation maximisation (EM) approach. It was although also noticed that the EM approach was far more sensitive to mixing matrix initialization and hence less robust. The Gibbs sampler described in this paper was implemented as part of the review of this paper. Although the implementation was problematic due to the inherent indeterminacy on gain, the approach provided good results to noisy linear mixtures of sources.

The recent paper [14] describes a hybrid approach where a general linear instantaneous model, which might be noisy and underdetermined, is tackled using a Student t distribution to model the source prior. Both the Gibbs sampler and VB method were studied as a method for inference. Both these methods are algorithmically similar but by employing a hybrid method the authors managed to combine the speed of the variational approach with the robustness of the Gibbs sampler [14].

1.1.6 Audio classification

This section contains a brief review of the methods typically employed in the field of audio classification. Classification features and classifiers are covered somewhat independently since, in many cases, they can be combined according to the particular application.

A variety of different features have been proposed for audio classification but features are mainly centered around the distinction between music, speech, silence and background noise. Zero-crossing rate (ZCR) has been reported as a useful feature for a variety of audio signals although again mainly used in speech/music classification algorithms[80]. ZCR is defined as the number of time-domain zero-crossings within a frame in the signal. Typically this number is divided by the number of samples within this frame. This feature is a popular feature used in music information retrieval and speech recognition[104][105][**?**]. ZCR is also a feature commonly used in voice activity detectors (VAD) along with energy based detectors[103].

Spectral flux (SF) compares the power spectrum of neighbouring audio frames in an effort to measure how quickly the spectrum fluctuates. Typically this is done by calculating the Euclidian distance between two normalised spectral coefficients[142].

The last audio feature presented here is the noise frame ration (NFR) metric. This features classifies frames as being noise frames if the local peak of their normalised correlation function reaches a certain threshold. The NFR is then the ratio between noisy frames and non-noisy frames in the audio segment[80][51].

Many other features have been proposed and used successfully throughout the literature. Other features include: An exhaustive list can be found here [105] and [**?**].

In the area of audio classification a couple of classifiers are commonly used. The Gaussian mixture model (GMM) classifier assumes that each class pdf is comprised of a specific number of multidimensional Gaussian distributions. Iterative expectation maximization (EM) can now be used to estimate parameters of the each Gaussian component and the mixture weights[127]. K nearest neighbors (K-NN) is a non-parametric classifier where samples are classified according to the majority of its K nearest neighbors[79]. Both of these classifiers have performed similarly to humans in genre classification applications[127] and in speech/music classification/segmentation applications K-NN has given total accuracy rates over 96% using a variety of features such as ZCR, SF and NFR among others[80]. Hidden Markov model (HMM) classifiers have also been employed where the HMM is trained using maximum likelihood (ML) based on a training sequence[64][147][10]. A classification algorithm employed in the context of HMM was the Viterbi algorithm which is covered in more detail later in this chapter[64][10]. A comprehensive introduction to various pattern recognition methods can be found in [23].

A study of 143 classification features for general audio data found that MFCC and linear prediction coefficients (LPC) provided “much better classification accuracy” compared to spectral features[74]. This was concluded irrespective of the number of audio categories desired.

1.1.7 Principal component analysis (PCA)

The central idea of PCA is to reduce the number of dimensions of a set of data consisting of possibly correlated variables to a smaller number of uncorrelated variables. This operation is achieved by transforming to a set of uncorrelated variables, principal components (PCs), which are ordered so the first PC retain most of the variability and each succeeding component accounts for as much of the remaining variability as possible.

Take $\mathbf{x}$ to be a vector of p random variables where the variance of the variables and the correlation or covariance between them are of interest. The approach taken by PCA is to look for a few ($<< p$) derived variables which preserves most of the information given by these variances and covariances or correlations. The initial step is to look for a linear function $\boldsymbol{\alpha}_1^T\mathbf{x}$ of $\mathbf{x}$ which maximizes the variance, where $\boldsymbol{\alpha}_1^T = [\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}]$ so that:

$$\boldsymbol{\alpha}_1^T\mathbf{x} = \alpha_{11}x_1 + \alpha_{11}x_2 + \ldots \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j, \tag{1.6}$$

is the first PC where ^{T} denotes a transposed matrix. Now, find the linear function $\boldsymbol{\alpha}_2^T\mathbf{x}$ of $\mathbf{x}$ uncorrelated with $\boldsymbol{\alpha}_1^T\mathbf{x}$ which again maximizes the variance, and so on, so that the kth PC $\boldsymbol{\alpha}_k^T\mathbf{x}$ maximizes the variances while being uncorrelated to $\boldsymbol{\alpha}_1^T\mathbf{x}, \boldsymbol{\alpha}_2^T\mathbf{x}, \ldots, \boldsymbol{\alpha}_{k-1}^T\mathbf{x}$. For a reduction of dimensionality it is hoped that most of the variability of $\mathbf{x}$ is accounted for by q PCs where $q << p$ [52, chap. 1].

A more common transform into uncorrelated basis functions is the discrete Fourier transform (DFT) or FT which decomposes a signal into a set of uncorrelated or orthogonal (in the statistical sense) complex exponentials of the form $e^{i\omega t}$. The transform described above can therefore be seen as a generalization of the DFT for random processes, where the decomposed uncorrelated basis functions are real random signals rather than complex exponentials [120, chap. 4.6].

1.1.8 Simple PCA Examples

Figure 1.1 gives a plot of 30 observations of two highly correlated (non zero mean) variables x_1 and x_2. It is seen that although there are considerable variation in the two variables, it is also noticed that there is more variation in the direction x_2.

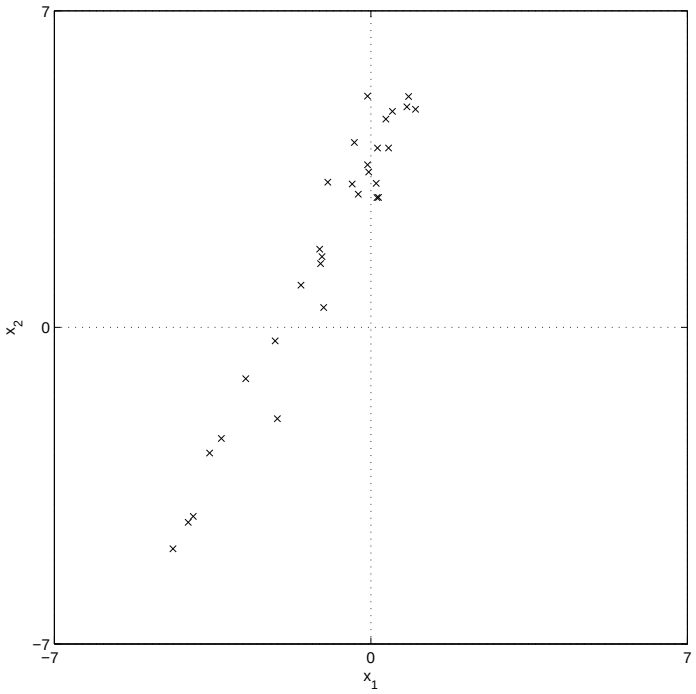


Figure 1.1: Plot of 30 observations of variables x_1 and x_2.

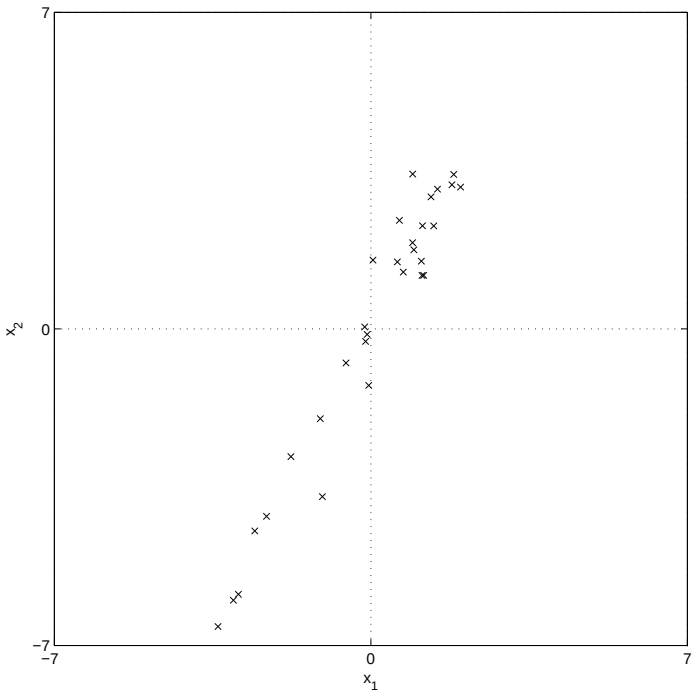


Figure 1.2: Plot of 30 observations of variables x_1 and x_2 with means subtracted.

It is also seen in Figure 1.1 that neither of the two variables are zero mean. To apply PCA the mean of both variables must be subtracted. Figure 1.2 shows the two variables with their means subtracted.

Since the process of PCA is closely related to the diagonalization of the correlation matrix [120, p. 174], a helpful interpretation of the PCA is the attempt to "rotate" the axes of Figure 1.2 to collect as much variability as possible in one dimension, rather than having it spread over 2 dimensions. The correlation matrix (correlation coefficients) for the two variables x_1 and x_2 is,

$$\text{Correlation matrix} = \text{corr}(x, y) = \begin{pmatrix} 1 & 0.97 \\ 0.97 & 1 \end{pmatrix}, \tag{1.7}$$

where the correlated nature of the two variables is clearly seen by the almost unitary off-diagonal terms. Since the covariance matrix $\mathbf{\Sigma}$ is square it is possible to look at the eigenvalues and eigenvectors of it, which are,

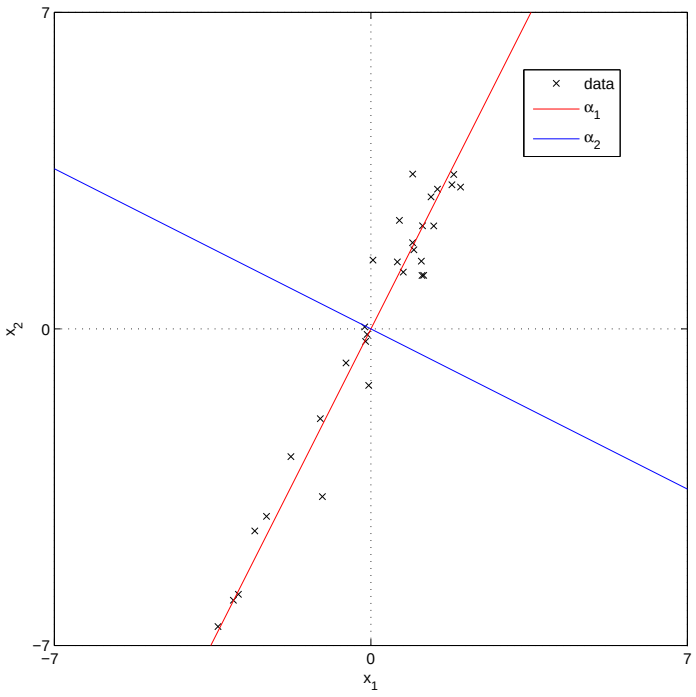


Figure 1.3: Plot of 30 observations of variables x_1 and x_2 with means subtracted, and eigenvectors of covariance matrix $\mathbf{\Sigma}$ plotted on top of the data. $\boldsymbol{\alpha}_k$ refers to the kth eigenvector.

$$\text{Eigenvalues} = \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 11.85 \\ 0.107 \end{pmatrix} \tag{1.8}$$

$$\text{Eigenvectors} = \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2) = \begin{pmatrix} 0.451 & -0.892 \\ 0.892 & 0.451 \end{pmatrix}. \tag{1.9}$$

The eigenvectors from equation (1.9) have been plotted on top of the data in Figure 1.3. These eigenvectors reveal information about patterns in the data, and as expected a clear correlation is found by the eigenvector $\boldsymbol{\alpha}_1$ between the variables x_1 and x_2. The second eigenvector $\boldsymbol{\alpha}_2$ gives the other, less important, pattern in the data.

The eigenvector with the highest corresponding eigenvalue is the PC. Ordering the eigenvectors in terms of descending eigenvalue gives the components in order of significance. It turns out that, for $k \in \{1,2,\ldots,p\}$ the kth PC is given by $z_k = \boldsymbol{\alpha}_k^T \mathbf{x}$,

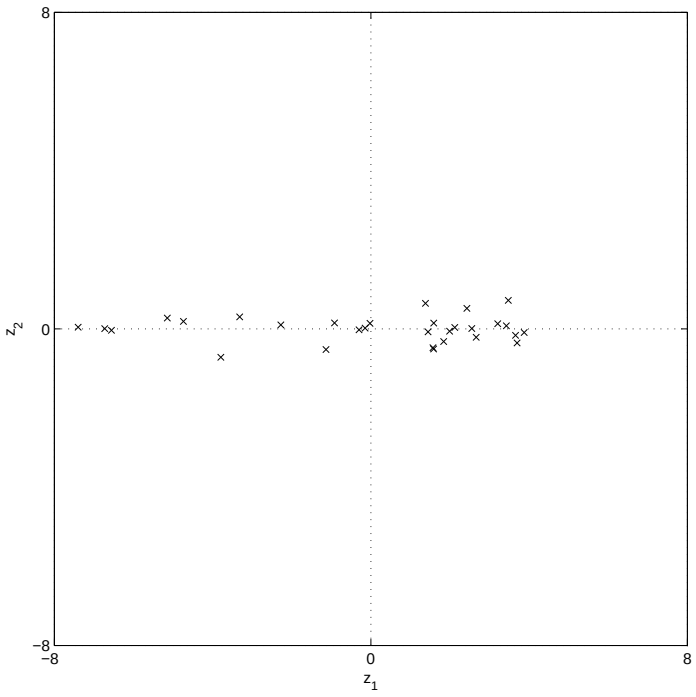


Figure 1.4: Plot of 30 observations of variables x_1 and x_2 transformed into z_1 and z_2. Almost all the information or variability is now in z_1.

where $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$ and where $\boldsymbol{\alpha}_k$ corresponds to the eigenvector with the kth largest eigenvalue [52, p. 2-3]. For a complete transformation of the data use the expression $\mathbf{z} = \hat{\boldsymbol{\alpha}}^T \mathbf{x}$ where $\hat{\boldsymbol{\alpha}}$ is the feature vector containing the number of eigenvectors desired q, so $\hat{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_q]$. Figure 1.3 shows the transformed data in terms of the transformed variables z_1 and z_2.

It is noticed that this transform, who's results are displayed in Figure 1.3, do not actually reduce the dimensionality of the data since $q = p$ and hence all information is preserved. By making $q < p$ the dimensionality is reduced and less significant data is discarded. Since it is noticed from equation (1.8) that λ_1 is much larger than λ_2, and hence $\boldsymbol{\alpha}_1$ is a far more significant component in the data than $\boldsymbol{\alpha}_2$, one could feasibly define the feature vector as $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_1$. Figure 1.5 shows the result of a transformation of the 30 observations for this scenario where $q = 1$. The data has been transformed back into x_1 and x_2 and the original means of the data has been reapplied.

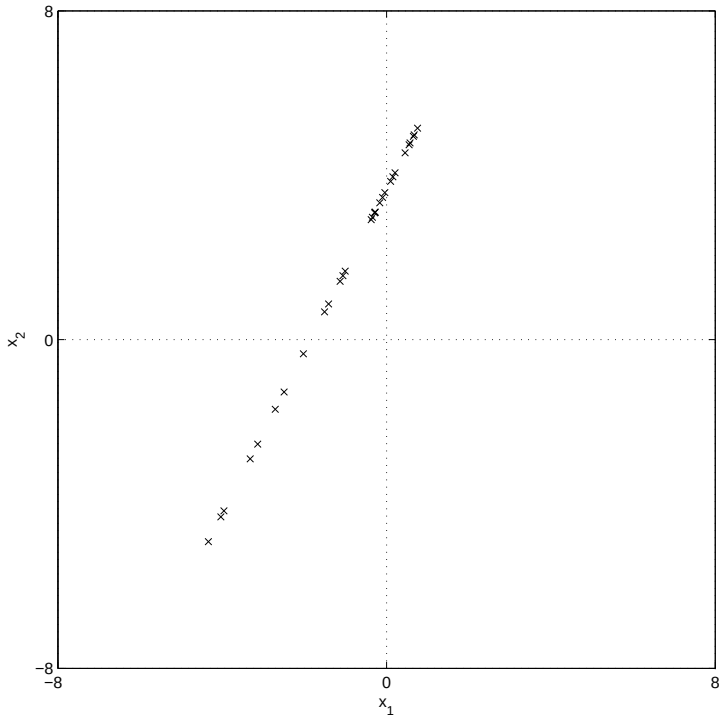


Figure 1.5: Plot of 30 observations of variables x_1 and x_2 transformed from only z_1 with original means added back in.

It is clearly seen how the data resembles that of Figure 1.1 although some of the variability has been lost in the dimensional reduction.

1.1.8.1 Eigenfaces

A popular application of PCA is the so called *eigenfaces*[109][126][151]. *Eigenfaces* are essentially eigenvectors derived from the covariance matrix of a high dimensional vector space or ensemble of individual pictures. The idea of *eigenpictures* was originally developed by [109] were the authors repeatedly draw parallels between how humans recognize facial features and their own computational method. Although the *eigenpictures* were not formerly used for facial recognition until [126], where the term *eigenfaces* was first coined, [109] hypothesises that perhaps humans compartmentalize faces and treat different features individually. This assertion is rooted first in the observation that humans are able to store and recognize enormous numbers of faces and secondly, that since recognition is apparently instantaneous it is conceivable that we do it by some efficient method, possibly similar to low dimensional methods. It is for example noted that "(...) fewer than 100 *eigenpictures* are necessary to fit a picture." and "fewer than 100 dimensions are needed to provide likeness.".

The method used in [109] uses photographs of 115 male undergraduates faces. These pictures were then digitized to a 128 by 128 pixel image and manually aligned. Figure 1.6 shows a figure of 3 cropped images. The left image is the average face based on the ensemble, the middle is a sample face and right image that sample faces departure from the ensemble average or the samples *caricature*. All pictures presented in this section are reproduced from the original paper [109] where none of the pictures have been filtered to eliminate the high frequencies produced by digitization.

The covariance matrix of the aligned ensemble is now calculated and all the *eigenpictures* determined. Figure 1.7 shows the first 8 of these *eigenpictures* going from the top left frame, moving right and ending on the bottom right frame with the 8th *eigenpicture*.

Figure 1.8 shows the approximate reproduction of the sample face from Figure 1.6 reproduced using 10, 20, 30 and 40 components.

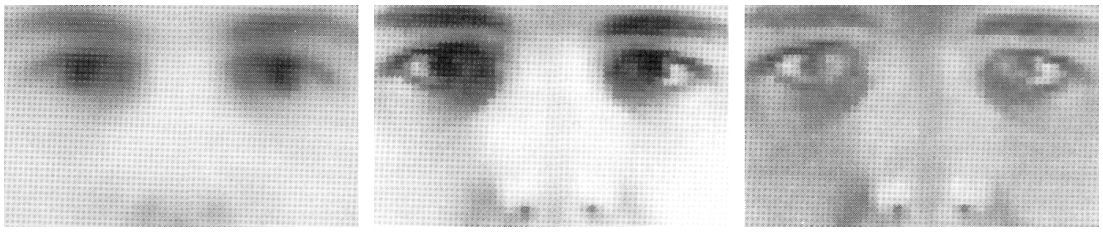


Figure 1.6: Cropped faces: left, the average; middle, a sample face; right, its caricature.

Figure 1.7: First 8 *eigenpictures* starting at upper left, moving to the right, and ending at lower right.

Figure 1.8: Approximation of the original picture (middle picture of Figure 1.6) using 10, 20, 30 and 40 *eigenpictures*.

1.1.8.2 In the literature

As mentioned previously, one of the main applications of PCA is dimensionality reduction which has rendered it useful in a variety of compression applications[133][3][151]. Within the image compression community [133] the focus on PCs has been useful in dimensionality reduction, and more specifically it has been widely applied as a way of representing animations of 3D geometric shapes. [3] presented an "easy and adaptive lossy compression" algorithm which provided compression of animation sequences with a factor of 1:100 accepting loss in animation accuracy. Another approach to animation compression is presented in [56] where PCA is combined with Linear Prediction Coding (LPC) to individually focus on spatial and the resulting temporal components respectively.

PCA has been adopted as a way to manage the enormous amounts of data produced by microarray experiments in genetics[100][101]. It was found that most of the observable variation in the data could be accounted for in just two components with a rapidly falling eigenvalues[100]. In other words, PCA reduces the data to two data sets which carry capture most of the information[100]. Later research has highlighted some advantages in ICA over PCA for this application[101]. Both approaches have shown to be robust to significant noise levels, while ICA showed some domain

specific feature extraction advantageous.

The MPEG-7 standard[12] has adopted a generalized sound recognition framework where decorrelated demension-reduced features, called audio spectrum project (ASP), are used to train hidden Markov models (HMM) for classification[61]. These features are regularly extracted from a basis decomposition via ICA or PCA[12][63][61]. ASP features for sound recognition has been compared to Mel-frequency cepstrum (MFCC) features and been found to be slower and have worse performance[62] while similar MPEG-7 related facial recognition tasks have yielded good performance but again with a significant computational cost for the feature extraction step[151].

Although PCA is optimal in approximating the input data in the mean-squared error sense, the representation that it provides is often not the most meaningful in terms of real world data and in terms of describing the fundamental properties of the data. Since the PCA describes the data in an orthonormal basis, purely in order of the second-order statistics (covariance) of the input data[95], which in effect means that PCA is networks are only able to realize linear input-output mappings[54]. In the field of neural networks there has been an interest in the development of non-linear PCA methods that take higher-order statistics into account.

Although not specifically a PCA approach, [44] presents an approach to general nonlinear generative model for non-linear factor analysis which can form the basis for many non-linear implementations of latent variable models such as a non-linear generalization of PCA or even ICA. This general model is based on the variational Bayesian framework, which forms a solid foundation for non-linear modeling [44]. This model can be implemented into a linear ICA framework for nonlinear blind source separation (BSS), and especially [129] found that the variational Bayesian method provided "useful" results for difficult nonlinear problems. [69] introduced nonlinear counterparts of PCA and ICA where the generative mapping from sources to data is not restricted to being linear. Here this mapping is modeled by a multi-layer perceptron (MLP) network and the distributions of source signals are modeled by Gaussians. The general form of the models discussed in this paper are of the form:

$$\mathbf{x}(t) = f(\mathbf{s}(t)) + \mathbf{n}(t), \tag{1.10}$$

where $\mathbf{x}(t)$ are the observations at time t, s(t) are the sources, $\mathbf{n}(t)$ the noise and $f()$ the function which maps the sources to the observation space. In [69] the authors compare their approach favorably to previously suggested models for representing data with nonlinear coordinate systems, and they especially focus on their methods applicability in high dimensional applications.

Very recent work has compared a range of multi-linear function factorisation techniques for feature extraction for classification applications[10]. The comparison was performed on a variety of data types including speaker recognition from spoken vowels. The article confirms previously presented results and emphasises the advantage and simplicity of the factorisation techniques in question in relation to HMMs and the associated training required. The authors found that canonical variates (CV) and PCA generally produced features with the greatest separability of the ones tested but neither reportedly performed well for the audio application. Specifically it was found in [10] that the PCA produced features of poor separability somewhat in contrast to [100]. This could potentially be due to the stacking, or lack thereof, employed by the authors. It is also noted that while PCA features, per definition, are ordered according to "principality", CV features require additional training[10].

1.1.9 Probabilistic PCA (PPCA)

One of the notable features of the above derived definition of the PCA is the lack of a probabilistic model for the observed data. [122] proposes a latent variable model for determining the principal axis of observed data which is closely related to factor analysis. This model utilizes a Maximum Likelihood (ML) estimator to estimate the parameters of the latent variable model. This model is commonly known as a factor analysis model where the relationship is linear:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{1.11}$$

in which the d-dimensional observations vector $\mathbf{t}$ is related to a corresponding q-dimensional vector of latent (or unobserved) variables $\mathbf{x}$. The $d \times q$ matrix $\mathbf{W}$ relates the two sets of variables, $\boldsymbol{\epsilon}$ is a zero-mean Gaussian noise process and the parameter $\boldsymbol{\mu}$ allows for model to have a non-zero mean. By defining $\mathbf{x} \sim \mathcal{N}_{\mathrm{x}}(\mathbf{0}, \mathbf{I})$ and

$\boldsymbol{\epsilon} \overset{i.i.d.}{\sim} \mathcal{N}_\epsilon(\mathbf{0}, \boldsymbol{\Psi})$, equation (1.11) induces a Gaussian distribution for the observations $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi})$. The parameters of this model can then be obtained through an iterative process using ML [122].

[70] presents an extended PPCA model based on the work of [122] termed Dual Probabilistic PCA (DPPCA). The DPPCA method can, through Gaussian processes, non-linearize the linear mappings from the embedded space and hence provide new probabilistic approach to visualizing and modeling of high dimensional data.

1.2 Detection

Any restoration of transient noise events presupposes complete knowledge of the position of the corruption. In practice this information is unknown *a priori* and a detection procedure must be employed to ascertain the timing of the corruptions. As noted in this section, a large number of different approaches to transient noise detection has been studied, and since there are as many types of transient noise as there are data that can be corrupted by it, the detection methods range from simple *ad hoc* filtering approaches to more complicated model based approaches.

In a reconstruction algorithm it is important to focus our attention on audible corruptions. The audibility of a corruption is not only a function of its amplitude or the energy that it represents, but also the context in which is sits. Psychoacoustically some corruptions may be rendered inaudible through masking effects such as the precedence effect, and hence any restoration effort is wasted or potentially damaging[88].

The simplest pulse detection algorithms exploit the relatively sparse nature of many audio, and in particular speech, signals above 9000 Hz in relation to impulsive noise which often exhibits a much smoother and wider frequency characteristics [116]. In [58][118] the authors preprocess their signals using a high-pass filter to target to the impulsive noise.

1.2.1 Median filter methods

A classic pulse noise detection (and restoration) scheme has involved a median filter[124][73][41][42][87][58]. Median filtering for signal smoothing, first published

in [124], has, according to [8], some important characteristics in that they reduce "spiky" noise while preserving jump discontinuities (edges). In [73] it is noted that the media filter has limited effect on non-impulsive noise and the authors propose a method for augmenting the median filter with a linear filter for added smoothing. This augmentation of the nonlinear media filter approach with a linear filtering approach has become a popular variation of median filtering process for impulsive noise detection and reduction spawning a variety of implementations [73][41][91][58][78]. In [59] it was also noted that the pulse detection algorithm that performed best was the median filter preprocessed by a linear filter. In recent years median based algorithms such as weighted median (WM) filters [150][141] and switching median filters [1][16][17][77] have seen a fair bit of attention although the focus of the implementations have almost exclusively been focused on image data.

The authors of [15] employ the SD-ROM (signal dependent rank order mean) algorithm, similar to the media filter methods, to evaluate the likelihood of each sample being corrupted based on the neighbouring samples. While this algorithm has shown great results in the removal of pulse noise in images [1] in [15] the method performs best for short, although frequent, noise pulses of the order of a single or a few samples.

Since transient noise events often exhibit a sudden fast change in the signal, one way to detect the onset of a noise event is to detect abrupt nonstationary changes in the dynamics of time series. In [31] the authors employ neural network predictors for this task, while [59] proposes an iterative discrete derivative method. In [59] the linear predictor and median hybrid method outperformed the derivative approach and the method described in [31] was never tested on real data but the requirement for training and the inherent detection deadzone does reduce the method's general applicability.

1.2.2 Autoregressive (AR) methods

While autoregressive (AR) methods have been used in other fields for detection and restoration of transient noise events [4] these methods were generally pioneered in the field of audio processing in [130][131][132]. AR methods are today the basis for

many pulse detection algorithms in audio applications[55][29][45][26][59][146][116]. The AR method proceeds by considering a sub-frame of the audio data x_t for $t = \{1,\ldots,N\}$. Assuming the data is drawn from a short-term stationary AR process:

$$x_t = \sum_{i=1}^{P} a_i x_{n-i} + e_t, \tag{1.12}$$

where e_t is the prediction error (or excitation signal) and $\mathbf{a} = \{a_1,\ldots,a_P\}$ is the AR coefficients of order P. The transient nature of the impulsive noise pulses will most likely lead to very large prediction errors if an attempt is made to predict its values with previous values of x_t. It follows that if an inverse AR filter is applied to an AR signal segment corrupted with transient noise events y_t, the prediction error $e_t = y_t - \sum_{i=1}^{P} a_i y_{n-i}$ is expected to be large when noise events are present while remaining low at other times[36].

In [132] the authors find that linear prediction systems, or AR process, are "adequate for modelling of speech signals whereas they can not model impulsive disturbances.". This realisation is used to separate out the residual of the LPC model effectively leaving the excitation noise in addition to the transient noise events, similarly to the pre-processing step of the favored approach in [59].

A noisy speech signal y_n can be modelled as an instantaneous mixture of a speech signal x_n and some some impulsive disturbance:

$$y_n = x_n + d_n, \tag{1.13}$$

and assuming that the speech signal can be modelled with a linear prediction model,

$$x_n = \sum_{k=1}^{P} a_k x_{n-k} + g e_n, \tag{1.14}$$

where $\mathbf{a}$ is the LPC parameters, e_n is the excitation signal and g is the linear prediction system's gain. The excitation signal can either be noise-like or a mixture of noise and some quasi periodic train of pulses. White noise is considered a good excitation signal for speech modelling[132].

Equation 1.13 can now be rewritten

$$y_n = \sum_{k=1}^{P} a_k x_{n-k} + e_n + d_n. \tag{1.15}$$

With an estimate for the LPC parameters $\hat{\mathbf{a}}$, the noisy signal y_n can be written as the noisy excitation signal v_n

$$\begin{aligned} v_n &= y_n - \sum_{k=1}^{P} \hat{a}_k y_{n-k} \\ &= x_n + d_h - \sum_{k=1}^{P} (a_k - \tilde{a}_k)(x_{n-k} + d_{n-k}), \end{aligned} \tag{1.16}$$

where a_k is the error in the LPC parameter vector estimate. Equation 1.16 can now be rewritten

$$v_n = e_n + d_h - \sum^{P} + k = 1 \hat{a}_k d_{n-k} + \sum_{k=1}^{P} \tilde{a}_k x_{n-k}. \tag{1.17}$$

The three elements that contribute to the noise in the excitation sequence estimate is therefore the impulsive disturbance d_n, the past noise samples p and the inflation in the variance of the the residual signal due to the error parameter estimate[132].

Since the transient noise pulses are transformed to a scaled version of the pulse response of the inverse LPC filter, and since experimental results conducted by [132] shows the amplitude of the excitation signal is in the order of 10^{-1} to 10^{-4} the detection task is greatly simplified. The authors of [37] note that the disadvantages with the approach outlined in [132] is the inability to detect small pulses in the presence of much larger disturbances as well as the introduction of distortion for certain signals. In [37] the pulse detection problem is put in a Bayesian framework and extended to non-Gaussian noise pulses.

An adaptation to the basic AR detection method in [131] uses a matched filter approach to detect transient noise events. The matched filter approach proceeds by considering the transient noise event as the signal and the AR data as the coloured additive noise[36]. An inherent problem with the matched filter is its dependence on training. Other methods in the literature model impulsive noise as non-Gaussian heavy-tailed distributions whereof the α-stable distribution is particularly popular [123][18]. According to [93] α-stable distributions are good for modeling many types of impulsive noise (including atmospheric and underwater acoustic noise). While the sub-Gaussians methods in [123][18] appear to perform well on certain kinds of

impulsive noise it is questionable whether they could perform in a real time application with high pulse variability and high sample rates.

Figure 1.9 and 1.10 compares the output of 4 basic detections schemes discussed up until now. The data presented in Figure 1.9(a) and Figure 1.10(a) is speech data corrupted with keyboard tapping noise sampled at 44.1 kHz. Each primary tapping pulse has been marked as "Ground Truth" although secondary pulses can also be seen. Comparing the matched filter (b) and AR prediction error (c) methods in Figures 1.9 and 1.10 it is noted that the matched filter produces a more smeared response while, in some cases, picking out more subtle pulses[36], e.g. 3rd marked pulse in Figure 1.9. The median filtered (d) and the high pass-filtered (HP) (e) signal produce nearly identical results. These results are also superior in that they increase the margin between lowest correct detection and the highest false detection is larger than for the AR prediction error (c) and potentially the matched filter (b). It is also noted that both me median and HP filtered signals pick up a faint secondary pulse just beyond the 6th confirmed pulse in Figure 1.9. For the matched filter (b) response this pulse is obscured due to temporal smearing. As will be discussed at a later point this secondary detection is typically associated with the lifting of a keyboard key and should therefor not be considered as a potential false detection.

Warped linear prediction (WLP) is another adaptation of the AR method [26]. The basic area of *warped* DSP was first introduced in [96] and later formalized in a predictive framework in [115] and as a recursive filter[114]. WLP has since been applied successfully to several audio applications [55][45] and more specifically used as a basis for pulse detection [29][26].

The basic concept of warped filters can be explained by considering a standard FIR-like structure, but rather than applying the standard unit delay z^{-1} the warped filter applies a new delay element $D(z)$ so that each new delay is frequency dependant (dispersive). In practice this means that the design of the warped filters are based on any pair of functions, $\tilde{z} = f(z)$ and $z = g(\tilde{z})$, so that $f(\cdot)$ and $g(\cdot)$ are one-to-one mappings of the unit circle onto itself, and $z = g\big(f(z)\big)$ [55]. Bilinear conformal mapping[9] conforms to the requirements and corresponds to the first order all-pass filter

$$\tilde{z}^{-1} = D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \tag{1.18}$$

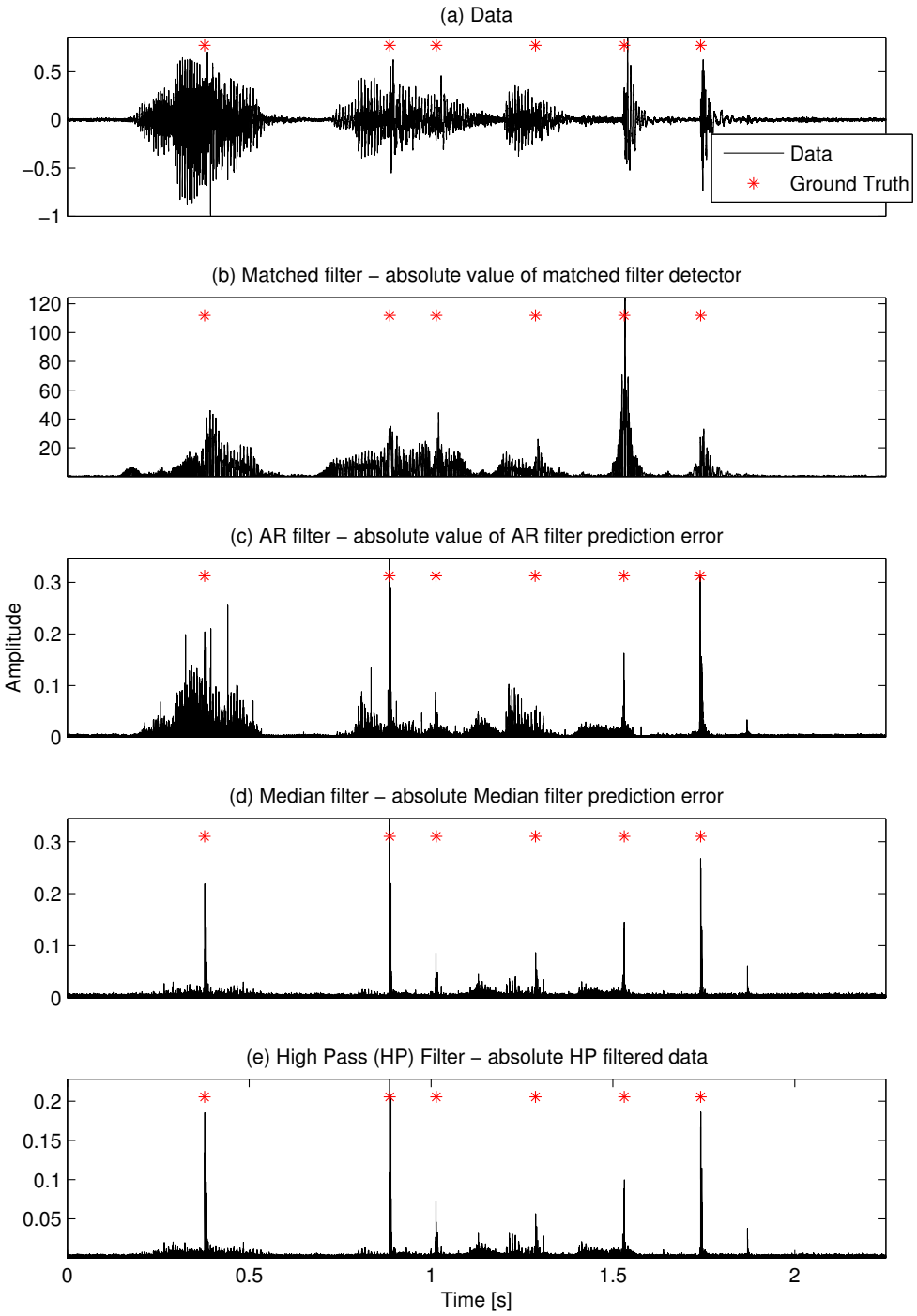


Figure 1.9: Pulse detection comparison, (a) Speech data with 6 primary keystroke pulses sampled at 44.1 kHz, (b) absolute value matched filter detector output, (c) absolute value of AR filter prediction error output, (d) absolute value of median filter prediction error, and (e) absolute value of high pass filtered data with crossover frequency 9.6 kHz.

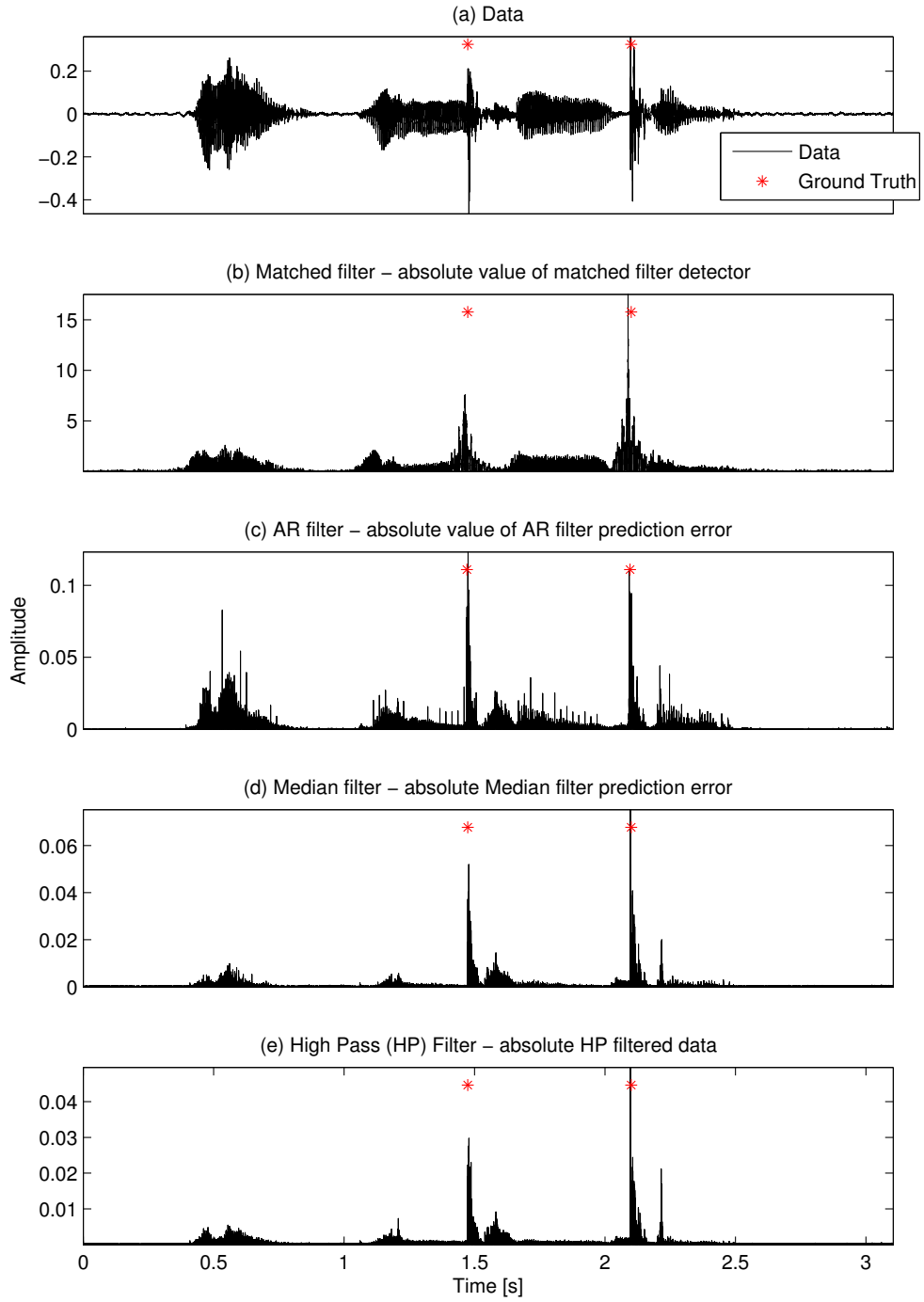


Figure 1.10: Pulse detection comparison, (a) Speech data with 2 primary keystroke pulses sampled at 44.1 kHz, (b) absolute value matched filter detector output, (c) absolute value of AR filter prediction error output, (d) absolute value of median filter prediction error, and (e) absolute value of high pass filtered data with crossover frequency 9.6 kHz.

where $\lambda, -1 < \lambda < 1$, is a warping parameter which, if chosen appropriately[55], yields a good match to the psychoacoustic Bark scale[113].

Warped digital filters have a range of advantages in that they can be designed to model the human auditory system as well as other physical systems[55]. In [26] the authors note that for auditory models the warping factor λ tend to be positive while for click detection negative warping factors appeared to perform best. It was also noted that pulse detection using WLP came at a computational cost and was therefore not suited for real time implementations. As in [36] the authors of [26] only considered pulses with duration of $< 1ms$, for which the WLP based method performs well. This is largely believed to be caused by spectral characteristics also exploited in [58][118], although the signals in [26] were not exclusively corrupted speech data and can therefore not be assumed to be spectrally sparse at high frequencies.

Warped-based methods evaluated via objective measures have although been found to only be advantageous for lower model orders while only being as good as conventional schemes otherwise[30][28]. Furthermore it is reported that warped-based methods increase the number of floating-point operations by around 77%[30].

1.2.3 Frequency methods

Others have attempted to use the short time Fourier transform (STFT) as a basis for detection [20][116][117] causing problems such as loss of temporal resolution of detections at moderate frame sizes, loss of spectral resolution for smaller frame size and computational inefficiency using extensive overlapping of frames. The authors of [116] propose an algorithm for detection of keystroke noise on laptop computers and recognise the temporal and spectral variability in the noise pulses causes methods based on noise models and stationarity assumptions to perform poorly. Instead the authors propose to exploit the "smoothness in speech signals present across time" and the relative spectral sparsity of speech signals compared to keystroke noise pulses with a simple linear predictive model across each frequency bin. Their model assumes that

$$S(k,t) = \sum_{m=1}^{M} \alpha_{km} S(k, t-\tau_m) + V(k,t), \tag{1.19}$$

where, $S(k,t)$ represents the time-frequency component for k and t, spectral and time index respectively, $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_M\}$ defines the frames used, $\boldsymbol{\alpha}_k = \{\alpha_{k1}, \ldots, \alpha_{kM}\}$ define the weights used for the linear prediction, and $V(t,k)$ is some zero-mean Gaussian noise with variance σ^2_{tk}.

The authors of [116] proceed to calculate the joint probability assuming independent frequency frames and eventually the log-likelihood F_t will be

$$F_t = -\frac{1}{2}\sum_k \frac{1}{\sigma^2_{tk}}\left(S(k,t) - \sum_{m=1}^{M} \alpha_{km} S(k, t-\tau_m)\right)^2 + C_{tk} \tag{1.20}$$

where C_{tk} is a constant.

1.2.4 Hidden Markov model (HMM)

While some detection algorithms quite simply base detections on magnitudes of some parameter[116][117] in some situations it may be advantageous to attempt to model, probabilistically, the evolution of a sequence of hidden states. For this application the hidden Markov model (HMM) is considered effective[98][148].

It is beyond the scope of this work to give a comprehensive introduction to HMM but particularly this source [98] is a good introduction to HMM in the field of speech processing.

The Viterbi algorithm, first proposed in 1967[138], is a recursive optimal solution to the problem of estimating the state sequence of a discrete-time finite-state Markov process[34]. Originally developed as a method for decoding convolutional codes from language identification [89]

Given an observation of a sequence $y \in \{y_1, \ldots, y_K\}$ the Viterbi algorithm's goal is to find the most probable sequence of states $S \in \{S_0, \ldots, S_K\}$ given these observations assuming the successive Markov state probabilities $Pr(S_{k-1} \rightarrow S_k)$ as well as the output probabilities $p(y_k | S_{k-1} \rightarrow S_k)$ are mutually independent for k. The likelihood function for the path from $k = 1$ to $k = K$ is given by

$$L = \prod_{k=1}^{K} Pr(S_{k-1} \rightarrow S_k)\, p(y_k | S_{k-1} \rightarrow S_k). \tag{1.21}$$

Typically the the logarithm is here considered due to computational concerns:

$$\log(L) = \sum_{k=1}^{K} m(y_k; S_{k-1}, S_k), \tag{1.22}$$

where m is the *branch metric* between two states S_{k-1}, S_k defined as

$$m(y_k; S_{k-1}, S_k) = \log(Pr(S_{k-1} \to S_k)) + \log\big(p(y_k | S_{k-1} \to S_k)\big). \tag{1.23}$$

The maximum *state metric* $M_K(S^i)$ over all paths leading from the origin to the ith state and Kth node S_K^i is defined as

$$M_K(S^i) = \max\left(\sum_{k=1}^{k-1} m(y_k; S_{k-1}, S_k) + m(y_K; S_{K-1}, S_K^i)\right), \tag{1.24}$$

for all paths $S_0, \ldots, S_{K-1}$. To maximize this sum you can simply maximize the first $K-1$ terms for each state S_{K-1}^j at the (K-1)th node, and then maximize the sum of this and the Kth term over all states S_{K-1}. In other words:

$$\begin{aligned} M_K(S^i) &= \max_{S_{K-1}^j}\Big(M_{K-1}(S^i) + m(y_K; S_{K-1}^j, S_K^i)\Big), \end{aligned} \tag{1.25}$$

which is the expression at the heart of the Viterbi algorithm[139].

1.2.4.1 Time-Frequency processing

Since the object of interest in our detection efforts is inherently transient and therefore localised in time, it is a significant shortcoming of classic Fourier analysis that it provides no such temporal information. The basics of time-frequency processing is the correlation of a signal with a family of waveforms that are well concentrated in time as well as in frequency[82] also called *time-frequency atoms*[35]. The popular STFT used in numerous applications dates back to 1946 and the introduction of the windowed Fourier atoms to measure the "frequency variations" of sound. Given the real and symmetric window

$$g_{u,\xi}(t) = \mathrm{e}^{i\xi t} g(t-u), \tag{1.26}$$

where ξ is a modulation frequency and u is a translation and normalized $\|g\| = 1$ so that $\|g_{u,\xi}\| = 1$ for any $(u,\xi) \in \mathbb{R}^2$. The resulting windowed Fourier transform of $f \in \mathbf{L^2}(\mathbb{R})$ is

$$Sf(u,\xi) = \langle f, g_{u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t)g(t-u)e^{-i\xi t}dt, \tag{1.27}$$

which is also called the STFT since the window $g(t-u)$ has the effect of localising the Fourier integral in the region of $t = u$.

To evaluate the energy density P_S of the STFT, also called the *spectrogram*, the squared magnitude is computed:

$$P_S f(u,\xi) = |Sf(u,\xi)|^2 = \left| \int_{-\infty}^{+\infty} f(t)g(t-u)e^{-i\xi t}dt \right|^2. \tag{1.28}$$

The *spectrogram* of f is a measure of the energy in the time-frequency neighborhood of (u,ξ). This is also called the Heisenberg box of $g_{u,\xi}$ and is defined as a region in the time-frequency plane (t,ω) whose location and width depends entirely on the time-frequency spread of the window $g_{u,\xi}$ centered around (u,ξ)[82].

For the windowed Fourier transform the time spread σ_t and frequency spread σ_w are independent of u and ξ. For more details on the derivation of the time-frequency resolution of the windowed Fourier transform see Appendix A.1 on page 40. Therefore $g_{u,\xi}$ corresponds to a Heisenberg box of area $\sigma_t\sigma_\omega$ centered at (u,ξ) as seen in Figure 1.11[43]. The size of the box is constant and therefore independent of (u,ξ) meaning that the windowed Fourier transform has the same temporal and frequency resolution throughout the time-frequency plane[82].

In practice this gives rise to a trade-off between temporal and frequency resolution, illustrated in Figure 1.13, where two different temporal frame sizes are shown in relation to the resulting frequency resolution.

1.2.4.2 Wavelet decomposition

The Wavelet Transform can be seen as a generalization of the windowed Fourier transform in that it decomposes a signal over dilated and translated wavelets. A wavelet is simply a function $\phi \in \mathbf{L^2}(\mathbb{R})$ which is normalised $\|\phi\| = 1$, centered in the neighborhood of $t = 0$, and with zero average:

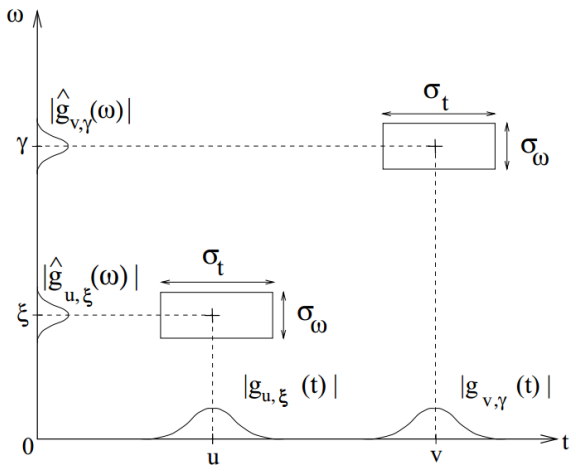


Figure 1.11: An example of the Heisenberg box illustration and the inherent time-frequency resolution trade-off for the STFT.

$$\int_{-\infty}^{+\infty} \psi(t)\,dt = 0. \tag{1.29}$$

Unlike the windowed Fourier transform the family of time-frequency wavelet atoms is translated by u as well as scaled by s:

$$\phi_{i,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right). \tag{1.30}$$

Now we have that the wavelet transform of $f \in \mathbf{L^2}(\mathbb{R})$ at time u and scale s is

$$Wf(u,s) = \langle f, \psi_{u,s}\rangle = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)dt. \tag{1.31}$$

The energy spread of a wavelet time-frequency atom $\phi_{u,s}$ corresponds to a Heisenberg box centered at $(u,\eta/s)$ where η is the center frequency of $\hat{\phi}$ the Fourier transform of ϕ, and $\hat{\phi_{u,s}}$ is the Fourier transform of ϕ dilated by $1/s$. For more details on the derivation of the time-frequency resolution of the wavelet transform see Appendix A.2 on page 41. The Heisenberg box remains of area $\sigma_t\sigma_\omega$ at all scales but it is now $s\sigma_t$ on the time axis and σ_ω/s along the frequency axis[82]. The temporal and frequency resolution is now dependent on s as illustrated in Figure 1.12

The main difference between classic Fourier analysis and Wavelet analysis is that Wavelets are localised in both frequency and time whereas the Fourier transform is

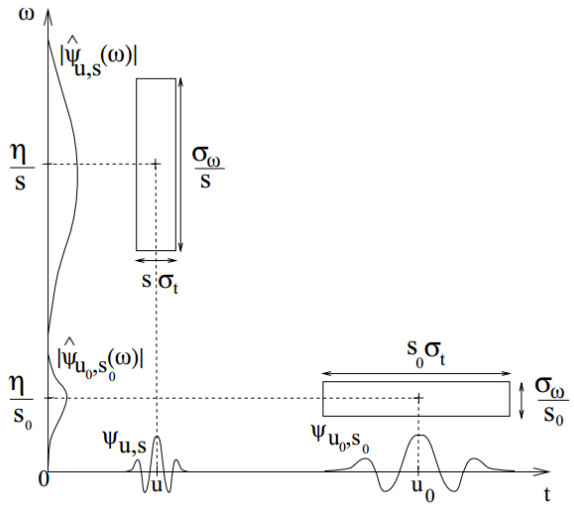


Figure 1.12: The Heisenberg boxes for the wavelet transform.

only localised in frequency. This leads to some significant advantages when considering data that is inherently transient. While the STFT does, to an extent, mimic the time localisation of the Wavelet Transform it does so at the cost of temporal resolution[82]. This inherent trade-off between temporal and frequency resolution is illustrated in Figure 1.13 where two different temporal frame sizes are shown in relation to the resulting frequency resolution.

Figure 1.14 shows a similar plot to Figure 1.13 but for the Wavelet transform instead. It should be noted that a Wavelet spectrum, equivalent to a traditional spec-

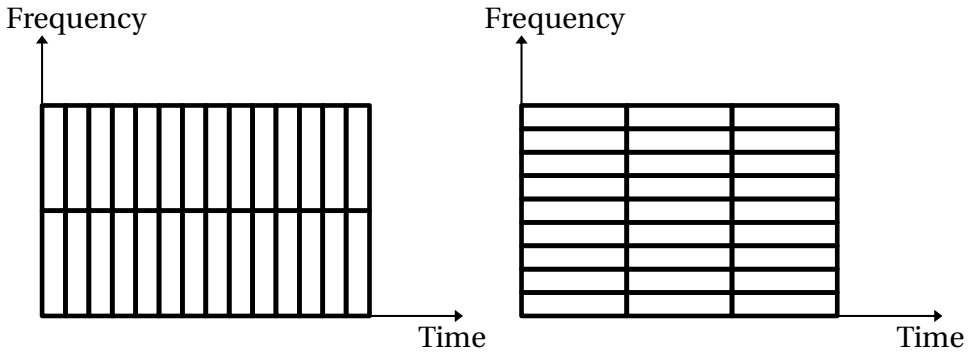


Figure 1.13: Heisenberg boxes of two wavelets. At larger scales the frequency resolution is increased by a decreased frequency support and the time spread is increased.

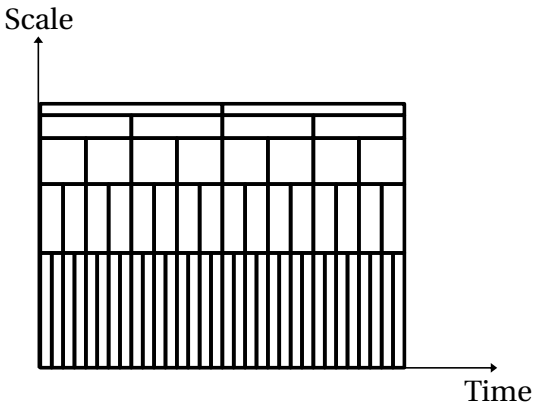


Figure 1.14: The scale-time relationship for the wavelet transform.

trogram, is traditionally plotted in the time-scale space, where scale can be seen as being inversely proportional to frequency.

1.2.4.3 Multi-resolution analysis with Filter Banks

Consider now the discrete power spectral density calculation, from Equation 1.28, for the signal $x(n)$:

$$S_{xx}(\omega) = \frac{1}{N}\left|\sum_{n=1}^{N} x(n)e^{-i\omega n}\right|^2. \tag{1.32}$$

For a given frequency ω_i, equation 1.32 can be rewritten as

$$S_{xx}(\omega_i) = \left\|\sum_{k=1}^{N} h_i(k)x(n-k)\right\|^2, \tag{1.33}$$

where

$$h_i(k) = w(k)e^{-i\omega_i n}, \tag{1.34}$$

and $w(k)$ is a window function. Considering $w(k)$ as a prototypical FIR lowpass filter, the collection of all $h_i(k)$s constitutes a bank of bandpass filters each centered on frequency ω_i[5]. This implementation is commonly referred to as the periodogram but it also highlights what sometimes is called the filter bank paradigm and is a particularly simple implementation of the DWT[82].

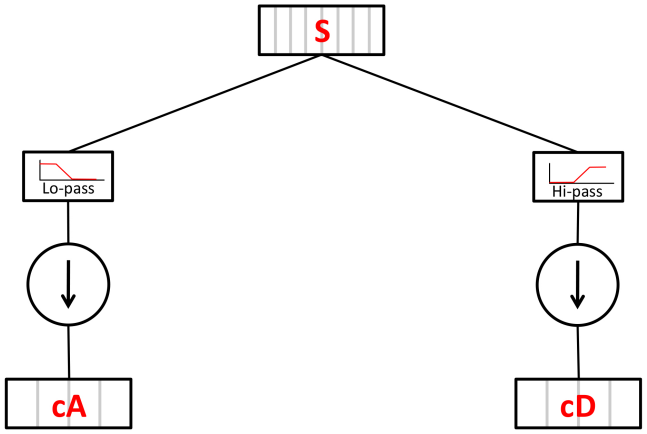


Figure 1.15: A single wavelet decomposition step.

Considering again the window function $w(k)$ from equation 1.34, the simplest window conceivable would be a simple rectangular window $w(k) = 1/N$. Naturally this would lead to significant side lobe leaks in the spectral domain and is therefore highly undesirable. Employing Hamming and Hann windows is a common approach to alleviate these issues. Another issue that arises with the periodogram is that its estimates are coarse with low precision and large variance[5]. A common method for alleviating this variance problem is by windowing the data first[75].

Work done on Multi Taper Spectrum Estimation (MTSE)[121] and later on multi-resolution theory[83][85] proves that conjugate mirror filter characterizes a wavelet and that cascading these filters will lead to a fast implementation of the discrete wavelet transform.

Figure 1.15 shows a diagrammatic representation of a single step of the DWT. The *Lo* and *Hi-pass* filters represent, respectively, the low and high pass conjugate mirror filters and the following step is a dyadic decimation (down-sampling) step. Since the spectrum of each filtered signal is effectively halved by the filters, the decimation step removes redundant information in accordance with Nyquist's theorem. The output of the low and high-pass filters are commonly referred to as the approximation and the detail coefficients respectively[82].

Figure 1.16 shows a three level wavelet decomposition.

The multi-resolution properties of the wavelet transform has made the Wavelet transform popular in a range of audio and speech applications where much of the

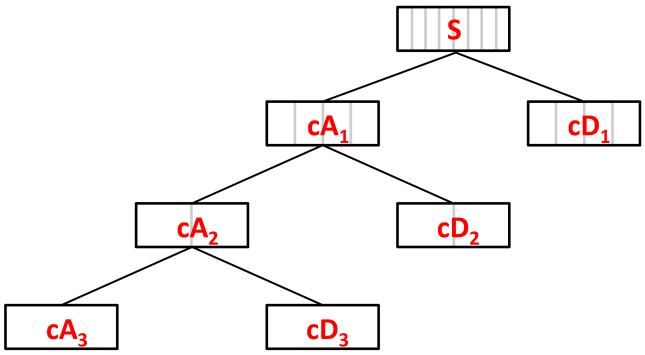


Figure 1.16: Three level wavelet decomposition tree.

information of interest is located in the lower frequency bands[108][20][68][6][128][153][76][94]. Equally the computational efficiency of the discrete wavelet transform is cited[53] as a significant advantage over the FFT with an $O(N)$ complexity compared to $O(N\log(N)$ for FFT[82]. In [53] the authors report "superior pitch detection performance" citing the Wavelet transforms computational efficiency, temporal resolution and its suitability of the pitch periods found in the analysed material. Speech applications in particular report good spectral estimation capabilities [46], good de-noising capabilities [21][106], and good compression performance [108][33]. Several authors propose wavelets used as a bases for audio classification applications[68][128][76] and report that a specific advantage of the wavelet basis is its compact representation compared to the time domain signal which decreases processing delay/cost[68].

An early wavelet based click detection algorithm proposes to detect clicks by analysing the wavelet coefficients of the signal for discontinuities[20]. The authors apply a neural network algorithm to robustly detect and classify pulses although the authors note that the required training stage is both "a complex and time-consuming procedure". While not explicitly discussed the authors appear to be targeting extremely short time pulses of the order of 1 or 2 samples, or what they call "parazite impulses".

More recently the wavelet bases has been used to detect impulsive noise in speech data by taking advantage of the relatively slow time-varying nature of speech and the Lipschitz regularity of the speech components[94]. For this application the scope of the corrupted samples appeared larger with suggested possible corruptions includ-

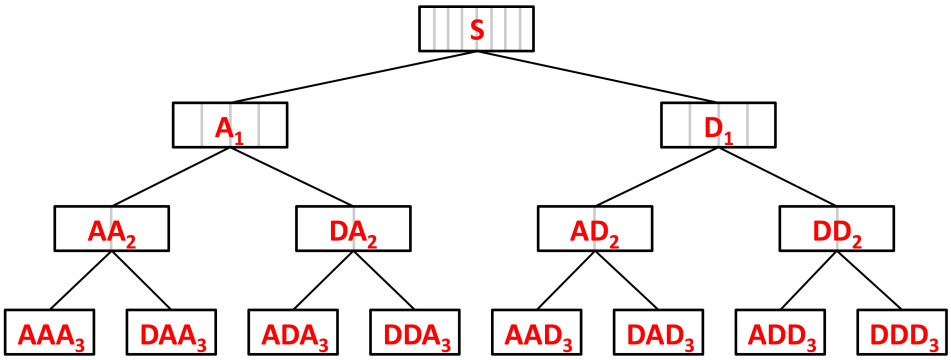


Figure 1.17: Three level wavelet packet decomposition tree.

ing gunshots, rain drops and keyboard typing noise. The author reported that the algorithm managed to reduce the rain drop noise in an instantaneous mixture of speech and rain noise. It was noted that this was the only noise application demonstrated and that it was never intended to completely remove the noise entirely. It is also noted that additional noise suppression algorithms seems to have been used as well as a speech enhancement algorithm. The extent to which this algorithm would reduce the nuisance of longer and louder pulses would have been interesting.

1.2.4.4 Wavelet Packet Transform (WPT)

The Wavelet Packet Transform (WPT) can be seen as a generalisation of the Wavelet transform in that it extends the link between multi-resolution approximations and wavelets. The WPT can also be seen as a natural extension of MSTE[121]. While the DWT decomposes only the approximation coefficients the WPT applies the decomposition step symmetrically throughout the tree as seen in Figure 1.17. Since the standard DWT algorithm is limited to wavelet bases that increase by a power of two towards the higher scales (low frequencies) it is possible that some other combination of bases could provide a better basis in some applications[19].

In a recent study comparing the WPT to other spectrum estimation techniques it was generally found that the WPT "operated well for all types of sources and its performances were comparable or at times even better than other existing approaches."[5]. In particular it was found that WPT performed well in reducing variance in the stop bands.

The authors of [40] developed a psychoacoustic model based around the WPT and found that not only was the WPT computationally more efficient than a Fourier transform-based alternative but it also provided a better spectral resolution. For psychoacoustic applications it was found that the generalization of the DWT yielded a better approximation of the critical bands employed[11][40] compared to similar DWT based models[108][153].

1.3 Restoration

The restoration of corrupted speech and audio samples by a localised degradation, also referred to as clicks or noise pulses, has inspired much research and many approaches. Corruptions can generally be seen to have varying severity. Clicks and pulses have traditionally been treated as completely corrupted or essentially missing samples transforming the restoration task into one of interpolation of missing samples[124][125][36].

An exhaustive review of the methods employed in the field up until the year of publishing in 1992 can be found in [134]. A summary of relevant methods and more recent methods are provided in this section. While the focus of this thesis is the interpolation of audio segments, as with the detection task, some methods or related approaches can be found in the field of image restoration of localised corruptions or other 1 and 2 dimensional implementations.

1.3.1 Nonlinear approaches

As noted in section 1.2.1, a classic pulse detection approach in both audio and image processing has been by median filtering. Equally, many of the approaches mentioned previously employ median based approaches for restoring the detected pulses[124][73][41][42][87][58][2]. In [7] it was shown that the optimal Order Statistics Filter (OSF) tends towards a median filter as the noise becomes increasingly impulsive, and that in particular the median filter is effective when the pulse length is less than half the median window size[2].

Much of the early work on nonlinear digital smoothing followed the original proposed algorithm by Tukey[124]. Following this publication the algorithm, in combi-

nation with linear filtering, was applied to speech signals and was found to perform reasonably[99]. The authors noted that the median filter alone, although successful in the preservation of sharp discontinuities not associated with noise, failed to provide sufficient smoothing and was therefore paired with a simple 3-point linear filter. In addition the authors described a double smoothing algorithm which attempted to isolate noisy segments and re-subtract them.

A similar early approach used to combat the effects of transmission errors in digital speech signals employed simple waveform-smoothing techniques and reported a "cleaner-sounding" speech in the presence of "fairly significant" error rates[50]. The result of this was unfortunately also a smearing of the speech component but the authors found that for high probability of errors this pulse-squelching operation was still perceptually desirable in spite of speech smearing. These results were backed up by a six person subjective study.

These methods all appear to have been applied indiscriminately to the data sequences which bore a cost to the signal fidelity.

Realizing the power of both linear and non-linear approaches to non-stationary noise detection and restoration, a range of *hybrid* approaches arose[92][73] [42]. A Modified Trimmed Mean (MTM) filter, was found to outperform a standard median filter in the task of preserving naturally sharp edged in signals[73]. The MTM filter quite simply replaces values of a frame with the average of a subframe selected around the outer frame's median value[73]. An FIR Median Hybrid (FMH) filter was designed with the aim of outperforming the standard median filter computationally. The FMH filter was found to perform similarly the the standard median filter[42]. Lastly the Adaptive Median Hybrid (AMH) filters, proposed in [92], employ multiple substructures which was considered "ideal for processing of signals with statistics which change abruptly". The theory is that at least one of the substructures will have adapted to a local change in the signal characteristics, whereas short non-stationary changes will be omitted. For the examples presented, this approach conserved sharp local changes while a comparable Least Means Squares (LMS) filter did not[92].

From a computational perspective median filters involve a sequence of searching and sorting steps and despite dedicated hardware for these operations these

procedures are generally computationally intensive[66]. A generalised mean implementation for black and white photos showed similar, and in some cases better, performance to the median filter while maintaining a distinct computational advantage[66].

In the field of image processing the median filter is still an integral part in most state-of-the-art impulsive noise filters[2]. Other popular non-linear methods employed have been the maximum-minimum method[149] and the peak-and-valley filter[145] and a modified maximum-minimum approach[2] which aims apply a maximum-minimum approach to selected samples only. All these non-linear approaches were found to have similar performance but with median filter bearing a higher computational cost[2].

A more direct and heuristic approach to waveform and audio restoration of longer gaps is found in waveform substitution[39] or "smart copying"[90]. An audio segment immediately prior to a corruption is used as a template for finding a segment earlier in the segment which matches it. It is assumed that what follows the best match for the template will be a good substitute for the corrupted segment. Given that speech signals "display quasi-stationary intervals"[39] this algorithm produces feasibly looking results given that the corruption does not straddle regions with high-energy voiced speed, low-level unvoiced speech or silence[39]. Waveform substitution has been confirmed through several objective listening tests to have "satisfactory" results[90]. In addition the method excels by being computationally simple[90].

An extension of the basic waveform substitution algorithm an additional pitch detection extension was proposed. If a pitch estimate is available the pitch period prior to the corruption may be reproduced throughout the corruption[39]. No direct results were presented to quantify the performance of this addition but a later publication ranks a pitch waveform replication algorithm as the producing the highest quality results based on subjective tests[143]. In this test the method was compared with a simple silence insertion and packet repetition algorithms suffering from obviously audible discontinuities[39] in addition to one and two-sided waveform substitution algorithms[39]. It was noted that while the pitch waveform replicator produced the highest quality results it was not the most computationally complex nor did it produce the longest delays of the tested algorithms[143].

A similar heuristic pitch based approach proposes to model speech as a series of spectral peak tracks that are “born” and “die” employing the short time Fourier transform (STFT)[81]. First proposed be [84] this method has since then been referred to as MQ representation (after the authors’ names) and been applied to the task of speech interpolation[81]. Many other authors have applied sinusoidal models for speech modeling in a variety of applications[36][135][144] which attempts in some cases to generalize the glottal excitation model in speech[84].

1.3.2 Linear approaches

In speech signals a popular and successful method of interpolation is the autoregressive (AR) approach[130][36][60]. An early approach using this method demonstrated that this method produced satisfactory results for specifically speech signals, based on both objective and subjective results, and furthermore established the feasibility of real-time implementation of such methods[49]. Independent development applied the methods to audio restoration applications[130][132].

Typically longer stretches of audio interpolation has been difficult applications for AR models[134][60], but modified AR based models have been applied to the pure extrapolation problem with some success to longer gaps[60]. The basic AR models, and later least squares (LS) AR extensions[36], have also been paired with synthesis filters designed to excite the AR model to achieve longer gap extrapolation with lower model orders[27]. This method achieves comparable perceptual quality to higher order AR models given significantly lower model complexity[27].

In general many authors have considered linear predictive coding (LPC) and AR models for audio and speech data[132][20][37]. A more comprehensive review of AR, LPC and warped-based methods were covered in section 1.2.2 from the perspective of modeling speech and audio in the detection application.

Appendix A

Time-frequency resolution details.

A.1 Windowed Fourier transform

To evaluate the energy density P_S of the STFT, also called the *spectrogram*, the squared magnitude is computed:

$$P_S f(u,\xi) = |Sf(u,\xi)|^2 = \left| \int_{-\infty}^{+\infty} f(t) g(t-u) \mathrm{e}^{-i\xi t} dt \right|^2. \tag{A.1}$$

The *spectrogram* of f is a measure of the energy in the time-frequency neighborhood of (u,ξ). This is also called the Heisenberg box of $g_{u,\xi}$ and is defined as a region in the time-frequency plane (t,ω) whose location and width depends entirely on the time-frequency spread of the window $g_{u,\xi}$ centered around (u,ξ)[82].

The time spread around u is independent of u and ξ:

$$\sigma_t^2 = \int_{-\infty}^{+\infty} (t-u)^2 |g_{u,\xi}(t)|^2 dt = \int_{-\infty}^{+\infty} t^2 |g(t)|^2 dt. \tag{A.2}$$

Since g is real and symmetric the Fourier transform of it $\hat{g}$ will also be real and symmetric.

$$\hat{g}_{u,\xi}(\omega) = \hat{g}(\omega-\xi) \exp\left[-i\,u(\omega-\xi)\right]. \tag{A.3}$$

The center frequency of the window $\hat{g}$ is now ξ and the frequency spread around it is:

$$\sigma_\omega^2 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (\omega-\xi)^2 |\hat{g}_{u,\xi}(\omega)| d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \omega^2 |\hat{g}(\omega)| d\omega. \tag{A.4}$$

For the windowed Fourier transform the time spread σ_t and frequency spread σ_w are independent of u and ξ. Therefore $g_{u,\xi}$ corresponds to a Heisenberg box of area $\sigma_t\sigma_\omega$ centered at (u,ξ) as seen in Figure 1.11[43]. The size of the box is constant and therefore independent of (u,ξ) meaning that the windowed Fourier transform has the same temporal and frequency resolution throughout the time-frequency plane[82].

A.2 Wavelet transform

The integral wavelet transform W of $f(t)$ is defined as:

$$Wf(u,s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt, \tag{A.5}$$

scales by s and translated by u[82].

Suppose that ϕ is centered at 0, so that $\phi_{u,s}$ is at $t = u$. The time-frequency spread of the wavelet atom $\phi_{u,s}$ determines the time-frequency resolution of the transform. Suppose that $v = \frac{t-u}{s}$ it can be verified that:

$$\int_{-\infty}^{+infty} (t-u)^2 |\psi_{u,s}|^2 dt = s^2 \sigma_t^2, \tag{A.6}$$

since

$$\sigma_t^2 = \int_{-\infty}^{+\infty} t^2 |\phi(t)|^2 dt. \tag{A.7}$$

At negative frequencies $\hat{\phi}(\omega)$ is zero, η, the center frequency of $\hat{\phi}$, is

$$\eta = \frac{1}{2\pi} \int_0^{+\infty} t^2 \omega |\hat{\phi}(\omega)|^2 d\omega. \tag{A.8}$$

The Fourier transform of $\phi_{u,s}$ can be calculated as $\hat{\phi}$ dilated by $1/s$, so that

$$\hat{\phi}_{u,s}(\omega) = \sqrt{s} \hat{\phi}(s\omega) \exp(-i\omega u). \tag{A.9}$$

Therefore η/s is the center frequency of $\hat{\phi}_{u,s}$ which has an energy spread of

$$\frac{1}{2\pi} \int_0^{+\infty} \left(\omega - \frac{\eta}{s} \right)^2 \left| \hat{\phi}_{u,s}(\omega) \right|^2 d\omega = \frac{\sigma_\omega^2}{s^2}, \tag{A.10}$$

where

$$\sigma_\omega^2 = \frac{1}{2\pi}\int_0^{+\infty} (\omega-\eta)^2 |\hat{\phi}(\omega)|^2 d\omega. \tag{A.11}$$

The energy spread of a wavelet time-frequency atom $\phi_{u,s}$ corresponds to a Heisenberg box centered at $(u,\eta/s)$ where η is the center frequency of $\hat{\phi}$ the Fourier transform of ϕ, and $\hat{\phi_{u,s}}$ is the Fourier transform of ϕ dilated by $1/s$. The Heisenberg box remains of area $\sigma_t\sigma_\omega$ at all scales but it is now $s\sigma_t$ on the time axis and σ_ω/s along the frequency axis[82]. The temporal and frequency resolution is now dependent on s as illustrated in Figure 1.12

References

[1] Eduardo Abreu, Michael Lightstone, Sanjit K Mitra, and Kaoru Arakawa. A new efficient approach for the removal of impulse noise from highly corrupted images. *Image Processing, IEEE Transactions on*, 5(6):1012–1025, 1996. doi: 10.1109/83.503916. URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=503916. 20

[2] Naif Alajlan, Mohamed Kamel, and Ed Jernigan. Detail preserving impulsive noise removal. *Signal Processing: Image Communication*, 19(10): 993 – 1003, 2004. ISSN 0923-5965. doi: http://dx.doi.org/10.1016/j.image.2004.08.003. URL http://www.sciencedirect.com/science/article/pii/S0923596504000761. 36, 38

[3] M. Alexa and W. Müller. Representing animations by principal components. *Computer Graphics Forum*, 19(3):411–418, 2000. doi: DOI:10.1111/1467-8659.00433. 16

[4] Kaoru Arakawa, Derek H. Fender, Hiroshi Harashima, Hiroshi Miyakawa, and Yoichi Saitoh. Separation of a nonstationary component from the eeg by a nonlinear digital filter. *Biomedical Engineering, IEEE Transactions on*, BME-33(7):724–726, 1986. ISSN 0018-9294. doi: 10.1109/TBME.1986.325767. 20

[5] Dyonisius Dony Ariananda, Madan Kumar Lakshmanan, and Homayoun Nikookar. An investigation of wavelet packet transform for spectrum estimation. *CoRR*, abs/1304.3795, 2013. 32, 33, 35

[6] Luiz WP Biscainho, Fvabio P Freeland, Paulo AA Esquef, and Paulo SR Diniz. Wavelet shrinkage denoising applied to real audio signals under perceptual

evaluation. In *EUPSICO 2000: European signal processing conference*, pages 2061–2064, 2000. 34

[7] A.C. Bovik, T.S. Huang, and Jr. Munson, D.C. A generalization of median filtering using linear combinations of order statistics. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 31(6):1342–1350, 1983. ISSN 0096-3518. doi: 10.1109/TASSP.1983.1164247. 36

[8] David R. Brillinger. John w. tukey's work on time series and spectrum analysis. *The Annals of Statistics*, 30(6):1595–1618, 2002. ISSN 00905364. URL http://www.jstor.org/stable/1558731. 20

[9] James Ward Brown and Ruel Vance Churchill. *Complex variables and applications*, volume 7. McGraw-Hill New York, 1996. 23

[10] M. Burke and J. Lasenby. Multilinear function factorisation for time series feature extraction. In *Digital Signal Processing (DSP), 2013 18th International Conference on*, pages 1–8, 2013. doi: 10.1109/ICDSP.2013.6622721. 7, 18

[11] B. Carnero and A. Drygajlo. Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms. *Signal Processing, IEEE Transactions on*, 47(6):1622–1635, 1999. ISSN 1053-587X. doi: 10.1109/78.765133. 36

[12] M. Casey. Mpeg-7 sound-recognition tools. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):737–747, 2001. ISSN 1051-8215. doi: 10.1109/76.927433. 17

[13] M. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. *in Proc. ICMC*, pages 154–161, 2000. 4

[14] A. Cemgil, C. Févotte, and S. Godsill. Variational and stochastic inference for bayesian source separation. *Digital Signal Processing*, 17(5):891 – 913, 2007. ISSN 1051-2004. doi: doi:10.1016/j.dsp.2007.03.008. Special Issue on Bayesian Source Separation. 5, 6

[15] C. Chandra, M.S. Moore, and S.K. Mitra. An efficient method for the removal of impulse noise from speech and audio signals. In *Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium on*, volume 4, pages 206 –208 vol.4, may-3 jun 1998. doi: 10.1109/ISCAS.1998.698795. 20

[16] Tao Chen and Hong Ren Wu. Impulse noise removal by multi-state median filtering. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 2183–2186 vol.4, 2000. doi: 10.1109/ICASSP.2000.859270. 20

[17] Tao Chen and Hong Ren Wu. Adaptive impulse detection using center-weighted median filters. *Signal Processing Letters, IEEE*, 8(1):1–3, 2001. doi: 10.1109/97.889633. 20

[18] M. J. Coates and E. E. Kuruoglu. Time-frequency-based detection in impulsive noise environments using alpha-stable noise models. *Signal Process.*, 82(12): 1917–1925, December 2002. ISSN 0165-1684. doi: 10.1016/S0165-1684(02)00319-5. URL http://dx.doi.org/10.1016/S0165-1684(02)00319-5. 22

[19] Ronald R Coifman, Yves Meyer, and Victor Wickerhauser. Wavelet analysis and signal processing. In *In Wavelets and their Applications*. Citeseer, 1992. 35

[20] A. Czyzewski. Some methods for detection and interpolation of impulsive distortions in old audio recordings. In *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, pages 139–142, 1995. doi: 10.1109/ASPAA.1995.482976. 26, 34, 39

[21] David L Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995. 34

[22] S. Douglas, M. Gupta, H. Sawada, and S. Makino. SpatioŰtemporal fastica algorithms for the blind separation of convolutive mixtures. *IEEE Transactions on ASLP*, 15(5):1511–1520, 2007. 3

[23] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2 edition, 2001. 7

[24] J. Eggink and G. Brown. Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In *Proc. ISMIR*, pages 125–131, 2003. 2

[25] D Ellis. *Prediction-Driven Computational Auditory Scene Analysis*. Ph.D. dissertationm, MIT, Cambridge, 1996. 1

[26] P. Esquef, M. Karjalainen, and V. Valimaki. Detection of clicks in audio signals using warped linear prediction. In *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, volume 2, pages 1085–1088 vol.2, 2002. doi: 10.1109/ICDSP.2002.1028279. 21, 23, 26

[27] P. A A Esquef and L. W P Biscainho. An efficient model-based multirate method for reconstruction of audio signals across long gaps. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1391–1400, 2006. ISSN 1558-7916. doi: 10.1109/TSA.2005.858018. 39

[28] Paulo A. A. Esquef, Luiz W. P. Biscainho, and Vesa Välimäki. An efficient algorithm for the restoration of audio signals corrupted with low-frequency pulses. *J. Audio Eng. Soc*, 51(6):502–517, 2003. URL http://www.acoustics.hut.fi/publications/papers/jaes-LP/. 26

[29] Paulo AA Esquef, Luiz WP Biscainho, Paulo SR Diniz, and Fhabio P Freeland. A double-threshold-based approach to impulsive noise detection in audio signals. In *Proc. X European Signal Processing Conf.(EUSIPCO 2000), Tampere, Finland*, pages 2041–2044, 2000. 21, 23

[30] Paulo AA Esquef, Vesa Välimäki, Kari Roth, and Ismo Kauppinen. Interpolation of long gaps in audio signals using the warped burgŠs method. In *Proc. 6th Int. Conf. on Digital Audio Effects (DAFx-03)*, pages 08–11. Citeseer, 2003. 26

[31] C.L. Fancourt and J.C. Principe. On the use of neural networks in the generalized likelihood ratio test for detecting abrupt changes in signals. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 2, pages 243–248 vol.2, 2000. doi: 10.1109/IJCNN.2000.857904. 20

[32] Cédric Févotte and Simon J. Godsill. A bayesian approach for blind separation of sparse sources. *in: Transactions on ASLP*, 14:2174–2188, 2006. 6

[33] E.-B. Fgee, W.J. Phillips, and W. Robertson. Comparing audio compression using wavelets with other audio compression schemes. In *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, volume 2, pages 698–701 vol.2, 1999. doi: 10.1109/CCECE.1999.808013. 34

[34] Jr. Forney, G.D. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, March 1973. ISSN 0018-9219. doi: 10.1109/PROC.1973.9030. 27

[35] D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, 1946. doi: 10.1049/ji-3-2.1946.0074. 28

[36] Simon J. Godsill and Peter J. W. Rayner. *Digital Audio Restoration*. Springer, 1998. ISBN 9783540762225. URL `http://www-sigproc.eng.cam.ac.uk/~sjg/springer/index.html`. 21, 22, 23, 26, 36, 39

[37] S.J. Godsill and P.J.W. Rayner. Statistical reconstruction and analysis of autoregressive signals in impulsive noise using the gibbs sampler. *Speech and Audio Processing, IEEE Transactions on*, 6(4):352 –372, jul 1998. ISSN 1063-6676. doi: 10.1109/89.701365. 22, 39

[38] D. Godsmark and G. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366, 1999. doi: doi:10.1016/S0167-6393(98)00082-X. 2

[39] D. Goodman, Gordon B. Lockhart, O. Wasem, and Wai-Choong Wong. Waveform substitution techniques for recovering missing speech segments in packet voice communications. *Acoustics, Speech and Signal Processing, IEEE*

Transactions on, 34(6):1440–1448, 1986. ISSN 0096-3518. doi: 10.1109/TASSP.1986.1164984. 38

[40] Xing He and Michael S. Scordilis. Psychoacoustic music analysis based on the discrete wavelet packet transform. *Res. Let. Signal Proc.*, 2008:4:1–4:5, January 2008. ISSN 1687-6911. doi: 10.1155/2008/346767. URL http://dx.doi.org/10.1155/2008/346767. 36

[41] P. Heinonen and Y. Neuvo. Smoothed median filters with fir substructures. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, volume 10, pages 49–52, 1985. doi: 10.1109/ICASSP.1985.1168502. 19, 20, 36

[42] P. Heinonen and Y. Neuvo. Fir-median hybrid filters. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(6):832–838, 1987. ISSN 0096-3518. doi: 10.1109/TASSP.1987.1165198. 19, 36, 37

[43] W. Heisenberg. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3-4):172–198, 1927. ISSN 0044-3328. doi: 10.1007/BF01397280. URL http://dx.doi.org/10.1007/BF01397280. 29, 41

[44] Antti Honkela and Harri Valpola. Unsupervised variational bayesian learning of nonlinear models. *Advances in Neural Information Processing Systems*, 17:593–600, 2005. 17

[45] Aki Härmä, Matti Karjalainen, Lauri Savioja, Vesa Välimäki, Unto K. Laine, and Jyri Huopaniemi. Frequency-warped signal processing for audio applications. *J. Audio Eng. Soc*, 48(11):1011–1031, 2000. URL http://www.aes.org/e-lib/browse.cfm?elib=12039. 21, 23

[46] Yi Hu and P.C. Loizou. Speech enhancement based on wavelet thresholding the multitaper spectrum. *Speech and Audio Processing, IEEE Transactions on*, 12(1):59–67, 2004. ISSN 1063-6676. doi: 10.1109/TSA.2003.819949. 34

[47] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, pages 626–634, 1999. 3

[48] Shiro Ikeda and Noboru Murata. A method of ica in time-frequency domain. In *in Proc. ICA*, pages 365–371, 1999. 3, 4

[49] A. Janssen, R. Veldhuis, and L. Vries. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(2):317–330, 1986. ISSN 0096-3518. doi: 10.1109/TASSP.1986.1164824. 39

[50] N.S. Jayant. Average- and median-based smoothing techniques for improving digital speech quality in the presence of transmission errors. *Communications, IEEE Transactions on*, 24(9):1043–1045, 1976. ISSN 0090-6778. doi: 10.1109/TCOM.1976.1093415. 37

[51] Hao Jiang, Tony Lin, and Hongjiang Zhang. Video segmentation with the support of audio segmentation and classification. In *Proc. IEEE ICME*, 2000. 7

[52] I. T. Jolliffe. *Principle Component Analysis*. Springer New York, 1986. doi: doi:10.1007/b98835. 8, 12

[53] S. Kadambe and G.F. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *Information Theory, IEEE Transactions on*, 38(2):917–924, 1992. ISSN 0018-9448. doi: 10.1109/18.119752. 34

[54] Juha Karhunen and Jyrki Joutsensalo. Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks*, 8(4):549 – 562, 1995. ISSN 0893-6080. doi: DOI:10.1016/0893-6080(94)00098-7. URL http://www.sciencedirect.com/science/article/B6T08-4031CR2-1S/2/1d83922db5bc73be2843e8b36b6166e1. 17

[55] M. Karjalainen, A. Harma, U.K. Laine, and Jyri Huopaniemi. Warped filters and their audio applications. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, pages 4 pp.–, 1997. doi: 10.1109/ASPAA.1997.625615. 21, 23, 26

[56] Zachi Karni and Craig Gotsman. Compression of soft-body animation sequences. *Computers & Graphics*, 28(1):25 – 34, 2004. ISSN 0097-8493. doi: DOI:10.1016/j.cag.2003.10.002. URL http://www.sciencedirect.com/science/article/B6TYG-4B8P413-1/2/321681a7a07f6a50138fdfa53d0935db. 16

[57] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of the Bayesian probability network to music scene analysis. *Computational auditory scene analysis*, pages 115–137, 1998. 2

[58] T. Kasparis and J. Lane. Adaptive scratch noise filtering. *Consumer Electronics, IEEE Transactions on*, 39(4):917–922, 1993. ISSN 0098-3063. doi: 10.1109/30.267417. 19, 20, 26, 36

[59] I. Kauppinen. Methods for detecting impulsive noise in speech and audio signals. In *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, volume 2, pages 967–970 vol.2, 2002. doi: 10.1109/ICDSP.2002.1028251. 20, 21

[60] Ismo Kauppinen and Kari Roth. Audio signal extrapolation–theory and applications. In *Proc. DAFx*, pages 105–110, 2002. 39

[61] H.G. Kim, N. Moreau, and T. Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley, 2006. ISBN 9780470093351. URL http://books.google.co.uk/books?id=eQM_Ip2nN8YC. 17

[62] Hyoung-Gook Kim and T. Sikora. Comparison of mpeg-7 audio spectrum projection features and mfcc applied to speaker recognition, sound classification and audio segmentation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 5, pages V–925–8 vol.5, 2004. doi: 10.1109/ICASSP.2004.1327263. 17

[63] Hyoung-Gook Kim, Edgar Berdahl, Nicolas Moreau, and Thomas Sikora. Speaker recognition using mpeg-7 descriptors. In *Proceedings of EUROSPEECH*, 2003. 17

[64] Don Kimber and Lynn Wilcox. Acoustic segmentation for audio browsers. *Computing Science and Statistics*, pages 295–304, 1997. 7

[65] T. Kinoshita, S. Sakai, and H. Tanaka. Musical sound source identification based on frequency component adaptation. In *Proc. IJCAI Worshop on CASA*, pages 18–24, 1999. 2

[66] A. Kundu, S.K. Mitra, and P.P. Vaidyanathan. Application of two-dimensional generalized mean filtering for removal of impulse noises from images. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(3):600–609, 1984. ISSN 0096-3518. doi: 10.1109/TASSP.1984.1164364. 38

[67] R.H. Lambert. *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. PhD thesis, Citeseer, 1996. 4

[68] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney. Classification of audio signals using statistical features on time and wavelet transform domains. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pages 3621–3624 vol.6, 1998. doi: 10.1109/ICASSP.1998.679665. 34

[69] H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multi-layer perceptrons. *Advances in independent component analysis*, pages 93–121, 2000. 17, 18

[70] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005. ISSN 1532-4435. 19

[71] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. doi: 10.1038/44565. 5

[72] T.W. Lee, A. Ziehe, R. Orglmeister, and TJ Sejnowski. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, volume 2. Citeseer, 1998. 4

[73] Yong-Hoon Lee and S.A. Kassam. Generalized median filtering and related nonlinear filtering techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(3):672–683, 1985. ISSN 0096-3518. doi: 10.1109/TASSP.1985.1164591. 19, 20, 36, 37

[74] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533 – 544, 2001. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/S0167-8655(00)00119-7. URL http://www.sciencedirect.com/science/article/pii/S0167865500001197. <ce:title>Image/Video Indexing and Retrieval</ce:title>. 7

[75] J.S. Lim and A.V. Oppenheim. *Advanced topics in signal processing*. Prentice-Hall signal processing series. Prentice-Hall, 1988. ISBN 9780130131294. URL http://books.google.co.uk/books?id=6uxSAAAAMAAJ. 33

[76] Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong, and Yukon Chang. Audio classification and categorization based on wavelets and support vector machine. *Speech and Audio Processing, IEEE Transactions on*, 13(5):644–651, 2005. ISSN 1063-6676. doi: 10.1109/TSA.2005.851880. 34

[77] Tzu-Chao Lin. A new adaptive center weighted median filter for suppressing impulsive noise in images. *Information Sciences*, 177(4):1073 – 1087, 2007. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2006.07.030. URL http://www.sciencedirect.com/science/article/pii/S0020025506002234. 20

[78] Jennifer C. Loveridge. Adaptiuve, hybrid median filter for temporal noise suppression, Jan. 1995. 20

[79] Lie Lu, Hao Jiang, and HongJiang Zhang. A robust audio classification and segmentation method. In *Proceedings of the Ninth ACM International Conference on Multimedia*, MULTIMEDIA '01, pages 203–211, New York, NY, USA, 2001. ACM. ISBN 1-58113-394-4. doi: 10.1145/500141.500173. URL http://doi.acm.org/10.1145/500141.500173. 7

[80] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10(7):504–516, 2002. ISSN 1063-6676. doi: 10.1109/TSA.2002.804546. 6, 7

[81] Robert C. Maher. A method for extrapolation of missing digital audio data. *J. Audio Eng. Soc*, 42(5):350–357, 1994. URL http://www.aes.org/e-lib/browse.cfm?elib=6946. 39

[82] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic press, 1999. 28, 29, 30, 31, 32, 33, 34, 40, 41, 42

[83] Stephane G Mallat. Multiresolution approximations and wavelet orthonormal bases of $l^2(r)$. *Transactions of the American Mathematical Society*, 315(1):69–87, 1989. 33

[84] R. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(4):744–754, 1986. ISSN 0096-3518. doi: 10.1109/TASSP.1986.1164910. 39

[85] Y. Meyer and D.H. Salinger. *Wavelets and Operators:*. Number v. 1 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995. ISBN 9780521458696. URL http://books.google.co.uk/books?id=y5L5HVlh3ngC. 33

[86] N. Mitianoudis and M. Davies. Audio source separation of convolutive mixtures. *IEEE Trans. Speech and Audio Processing*, 11(5):489–497, sep 2003. 3

[87] A. Mäkivirta, E. Koski, A. Kari, and T. Sukuvaara. The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. *Computer Methods and Programs in Biomedicine*, 34(2Ű3):139 – 144, 1991. ISSN 0169-2607. doi: http://dx.doi.org/10.1016/0169-2607(91)90039-V. URL http://www.sciencedirect.com/science/article/pii/016926079190039V. <ce:title>Knowledge-Based Systems in Medicine A Nordic Research & Development Programme</ce:title>. 19, 36

[88] Brian C. J. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition*. Academic Press, April 2003. ISBN 0125056281. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0125056281. 19

[89] T. Nagarajan and H.A. Murthy. Language identification using parallel syllable-like unit recognition. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1, pages I–401–4 vol.1, 2004. doi: 10.1109/ICASSP.2004.1326007. 27

[90] M. Niediwiecki and K. Cisowski. Smart copying-a new approach to reconstruction of audio signals. *Signal Processing, IEEE Transactions on*, 49(10): 2272–2282, 2001. ISSN 1053-587X. doi: 10.1109/78.950783. 38

[91] Ari Nieminen, P. Heinonen, and Y. Neuvo. A new class of detail-preserving filters for image processing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-9(1):74–90, 1987. ISSN 0162-8828. doi: 10.1109/TPAMI.1987.4767873. 20

[92] Ari Nieminen, P. Heinonen, and Y. Neuvo. Suppression and detection of impulse type interference using adaptive median hybrid filters. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, volume 12, pages 117–120, 1987. doi: 10.1109/ICASSP.1987.1169749. 37

[93] Chrysostomos L. Nikias and Min Shao. *Signal processing with alpha-stable distributions and applications*. Wiley-Interscience, New York, NY, USA, 1995. ISBN 0-471-10647-X. 22

[94] RC Nongpiur. Impulse noise removal in speech using wavelets. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1593–1596. IEEE, 2008. doi: 10.1109/ICASSP.2008.4517929. 34

[95] Erkki Oja and Juha Karhunen. Signal separation by nonlinear hebbian learning. *Computational intelligence: A dynamic system perspective*, pages 83–97, 1995. 17

[96] Alan V Oppenheim, Alan S Willsky, and Syed Hamid Nawab. *Signals and systems*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1983. 23

[97] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1):23–35, 1997. doi: 10.1016/S0169-7439(96)00044-5. 5

[98] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. ISSN 0018-9219. doi: 10.1109/5.18626. 27

[99] L. Rabiner, M. Sambur, and C. Schmidt. Applications of a nonlinear smoothing algorithm to speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(6):552–557, 1975. ISSN 0096-3518. doi: 10.1109/TASSP.1975.1162749. 37

[100] Soumya Raychaudhuri, Joshua M Stuart, and Russ B Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 455. NIH Public Access, 2000. 16, 18

[101] Samir A Saidi, Cathrine M Holland, David P Kreil, David JC MacKay, D Stephen Charnock-Jones, et al. Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*, 23(39):6677–6683, 2004. 16

[102] Y. Sakuraba and H. Okuno. Note recognition of polyphonic music by using timbre similarity and direction proximity. In *Proc. of ICMC*, pages 167–170, 2003. 2

[103] A. Sangwan, M.C. Chiranth, H.S. Jamadagni, R. Sah, R. Venkatesha Prasad, and V. Gaurav. Vad techniques for real-time speech transmission on the internet. In *High Speed Networks and Multimedia Communications 5th IEEE International Conference on*, pages 46–50, 2002. doi: 10.1109/HSNMC.2002.1032545. 6

[104] C. Saraceno and R. Leonardi. Identification of successive correlated camera shots using audio and video information. In *Image Processing, 1997. Proceedings., International Conference on*, volume 3, pages 166–169 vol.3, 1997. doi: 10.1109/ICIP.1997.632039. 6

[105] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334 vol.2, 1997. doi: 10.1109/ICASSP.1997.596192. 6, 7

[106] Jong-Won Seok and Keun sung Bae. Speech enhancement with reduction of noise components in the wavelet domain. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1323–1326 vol.2, 1997. doi: 10.1109/ICASSP.1997.596190. 34

[107] Yang Shao, Soundararajan Srinivasan, Zhaozhang Jin, and DeLiang Wang. A computational auditory scene analysis system for speech segregation and robust speech recognition. *Comput. Speech Lang.*, 24(1):77–93, 2010. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2008.03.004. 2

[108] D. Sinha and A.H. Tewfik. Low bit rate transparent audio compression using adapted wavelets. *Signal Processing, IEEE Transactions on*, 41(12):3463–3479, 1993. ISSN 1053-587X. doi: 10.1109/78.258086. 34, 36

[109] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, 1987. 14

[110] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34, 1998. doi: doi:10.1016/S0925-2312(98)00047-2. 3, 4

[111] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. *Lecture Notes in Computer Science*, pages 494–499, 2004. 5

[112] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003. 5

[113] J.O. Smith and J.S. Abel. The bark bilinear transform. In *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, pages 202–205, 1995. doi: 10.1109/ASPAA.1995.482991. 26

[114] K Steiglitz. A note on variable recursive digital filters. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(1):111–112, 1980. 23

[115] Hans Werner Strube. Linear prediction on a warped frequency scale. *The Journal of the Acoustical Society of America*, 68(4):1071–1076, 1980. doi: 10.1121/1.384992. URL http://link.aip.org/link/?JAS/68/1071/1. 23

[116] A. Subramanya, M.L. Seltzer, and A. Acero. Automatic removal of typed keystrokes from speech signals. *Signal Processing Letters, IEEE*, 14(5):363–366, may 2007. ISSN 1070-9908. doi: 10.1109/LSP.2006.888091. 19, 21, 26, 27

[117] Akihiko Sugiyama. Single-channel impact-noise suppression with no auxiliary information for its detection. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 127–130. IEEE, oct. 2007. doi: 10.1109/ASPAA.2007.4393053. 26, 27

[118] Kenichi Taura, Masahiro Tsuujishita, Masayuki Tsuji, and Masayuki Ishida. Impulse noise reducer detecting impulse noise from an audio signal, Sep. 2004. 19, 26

[119] F.J. Theis. Towards a general independent subspace analysis. *Advances in Neural Information Processing Systems*, 19:1361, 2007. 4

[120] Charles W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall, 1992. 8, 10

[121] D.J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982. ISSN 0018-9219. doi: 10.1109/PROC.1982.12433. 33, 35

[122] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (3):611–622, 1999. doi: DOI:10.1111/1467-9868.0019. 18, 19

[123] G.A. Tsihrintzis and C.L. Nikias. Data-adaptive algorithms for signal detection in sub-gaussian impulsive interference. *Signal Processing, IEEE Transactions on*, 45(7):1873–1878, 1997. ISSN 1053-587X. doi: 10.1109/78.599964. 22

[124] John W. Tukey. Nonlinear (nonsuperposable) methods for smoothing data. In *EASCON Conf. Rec.*, page p. 673, 1974. Also in CWJWT II (1985) 837?855. 19, 20, 36

[125] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley series in behavioral sciences. Addison-Wesley Publishing Company, 1977. ISBN 9780201076165. URL http://books.google.co.uk/books?id=UT9dAAAAIAAJ. 36

[126] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 14

[127] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on*, 10(5):293–302, 2002. ISSN 1063-6676. doi: 10.1109/TSA.2002.800560. 7

[128] George Tzanetakis, Georg Essl, and Perry Cook. Audio analysis using the discrete wavelet transform. In *Proc. Conf. in Acoustics and Music Theory Applications*, 2001. 34

[129] H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen. Nonlinear blind source separation by variational Bayesian learning. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 86:532–541, 2003. 17

[130] Saeed V. Vaseghi. *Algorithms for restoration of archived gramophone recordings*. PhD thesis, University of Cambridge, 1988. 20, 39

[131] S.V. Vaseghi and P. J W Rayner. A new application of adaptive filters for restoration of archived gramophone recordings. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 2548–2551 vol.5, 1988. doi: 10.1109/ICASSP.1988.197163. 20, 22

[132] S.V. Vaseghi and P.J.W. Rayner. Detection and suppression of impulsive noise in speech communication systems. *Communications, Speech and Vision, IEE Proceedings I*, 137(1):38 –46, feb. 1990. ISSN 0956-3776. 20, 21, 22, 39

[133] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear subspace analysis of image ensembles. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:93, 2003. ISSN 1063-6919. doi: http://doi.ieeecomputersociety.org/10.1109/CVPR.2003.1211457. 16

[134] Raymond Veldhuis. *Restoration of lost samples in digital signals*. Prentice-Hall, Inc., 1992. 36, 39

[135] P Vera-Candeas, N Ruiz-Reyes, J Curpián-Alonso, and M Rosa-Zurera. A new sinusoidal modelling approach for parametric speech and audio coding. In *Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the 3rd International Symposium on*, volume 1, pages 134–139. IEEE, 2003. 39

[136] E. Vincent. Musical source separation using time-frequency priors. *IEEE transactions on Audio, Speech and Language Processing*, 14(1):91–98, Jan 2006. doi: 10.1109/TSA.2005.860342. 2

[137] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007. 3, 5

[138] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2): 260–269, 1967. ISSN 0018-9448. doi: 10.1109/TIT.1967.1054010. 27

[139] A.J. Viterbi. A personal history of the viterbi algorithm. *Signal Processing Magazine, IEEE*, 23(4):120–142, 2006. ISSN 1053-5888. doi: 10.1109/MSP.2006.1657823. 28

[140] D. Wang and G. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley, 2006. 1, 2

[141] Gaihua Wang, Dehua Li, Weimin Pan, and Zhaoxiang Zang. Modified switching median filter for impulse noise removal. *Signal Process.*, 90(12):3213–3218, December 2010. ISSN 0165-1684. doi: 10.1016/j.sigpro.2010.05.026. URL http://dx.doi.org/10.1016/j.sigpro.2010.05.026. 20

[142] Wengen Wang, Xiaoqing Yu, Yun Hui Wang, and R. Swaminathan. Audio fingerprint based on spectral flux for audio retrieval. In *Audio, Language and Image Processing (ICALIP), 2012 International Conference on*, pages 1104–1107, 2012. doi: 10.1109/ICALIP.2012.6376781. 7

[143] O.J. Wasem, D.J. Goodman, C.A. Dvorak, and H.G. Page. The effect of waveform substitution on the quality of pcm packet communications. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(3):342–348, 1988. ISSN 0096-3518. doi: 10.1109/29.1530. 38

[144] Jeremy J Wells and Damian T Murphy. High accuracy frame-by-frame non-stationary sinusoidal modelling. In *Proceedings of the 9th International Conference on Digital Audio Effects, Montreal, Canada*, pages 253–258, 2006. 39

[145] P.S. Windyga. Fast impulsive noise removal. *Image Processing, IEEE Transactions on*, 10(1):173–179, 2001. ISSN 1057-7149. doi: 10.1109/83.892455. 38

[146] P.J. Wolfe and S.J. Godsill. Interpolation of missing data values for audio signal restoration using a gabor regression model. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 5, pages v/517 – v/520 Vol. 5, march 2005. doi: 10.1109/ICASSP.2005.1416354. 21

[147] Ziyou Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang. Comparing mfcc and mpeg-7 audio features for feature extraction, maximum likelihood hmm and entropic prior hmm for sports audio classification. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 5, pages V–628–31 vol.5, 2003. doi: 10.1109/ICASSP.2003.1200048. 7

[148] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. Hmm-based audio keyword generation. In Kiyoharu Aizawa, Yuichi Nakamura, and ShinŠichi Satoh, editors, *Advances in Multimedia Information Processing - PCM 2004*, volume 3333 of *Lecture Notes in Computer Science*, pages 566–574. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-23985-7. doi: 10.1007/978-3-540-30543-9_71. URL http://dx.doi.org/10.1007/978-3-540-30543-9_71. 27

[149] Y. Xu and E. M K Lai. Restoration of images contaminated by mixed gaussian and impulse noise using a recursive minimum-maximum method. *Vision, Image and Signal Processing, IEE Proceedings* -, 145(4):264–270, 1998. ISSN 1350-245X. doi: 10.1049/ip-vis:19981995. 38

[150] Lin Yin, Ruikang Yang, M. Gabbouj, and Y. Neuvo. Weighted median filters: a tutorial. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 43(3):157–192, 1996. ISSN 1057-7130. doi: 10.1109/82.486465. 20

[151] N. Zaeri, F. Mokhtarian, and A. Cherri. Extension of the mpeg-7 fourier feature descriptor for face recognition using pca. In *GCC Conference (GCC), 2006 IEEE*, pages 1–6, 2006. doi: 10.1109/IEEEGCC.2006.5686244. 14, 16, 17

[152] K. Zhang and L. Chan. Separating Convolutive Mixtures By Pairwise Mutual Information Minimization. *IEEE Signal Processing Letters*, 14(12):992–995, 2007. doi: 10.1109/LSP.2007.906224. 4

[153] M. Rosa Zurera, F. López Ferreras, M.P. Jarabo Amores, S. Maldonado Bascón, and N. Ruiz Reyes. A new algorithm for translating psycho-acoustic information to the wavelet domain. *Signal Processing*, 81(3):519 – 531, 2001. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/S0165-1684(00)00230-9. URL http://www.sciencedirect.com/science/article/pii/S0165168400002309. <ce:title>Special section on Digital Signal Processing for Multimedia</ce:title>. 34, 36