Selection of Relevant Features for Audio Classification tasks

(Philosophy of Doctoral)

Maria Markaki

Heraklion

October 2011

Department of Computer Science
University of Crete

Selection of Relevant Features for Audio Classification tasks

Submitted to the
Department of Computer Science
in partial fulfillment of the requirements for the degree of
Philosophical Doctoral

Maria Markaki

web: http://www.csd.uoc.gr/~mmarkaki/

e-mail: mmarkaki@csd.uoc.gr and
maria.g.markaki@gmail.com

October, 2011

Board
of inquiry:

Supervisor _____
Yannis Stylianou
Professor

Member _____
Petros Maragos
Professor

Member _____
Georgios Tziritas
Professor

Member _____
Vassilis Digalakis
Professor

Member _____
Alexandros Potamianos
Associate Professor

Member _____
Athanasios Mouchtaris
Assistant Professor

Member _____
Ioannis Tsamardinos
Assistant Professor

Heraklion, October, 2011

Abstract

Advances in time-frequency distributions and spectral analysis techniques (i.e., for the estimation of amplitude and/or frequency modulations) allow a better representation of non-stationary signals like speech, highlighting their fine structure and dynamics. Although such representations are very useful for analysis purposes, they complicate the classification tasks due to the large number of parameters extracted from the signal ("curse of dimensionality"). For such tasks, a significant dimensionality reduction is required.

In this thesis, the problem of dimensionality reduction of these time/frequency-frequency representations is studied; selection criteria of the optimal parameters are suggested, based on their relevance to a given classification task. Relevance is defined based on mutual information. First, using tools from multilinear algebra, such as High Order SVD, the initial dimensions and the noise components of the representation are reduced. Then, feature selection proceeds based on maximum relevance criterion. It is shown that the suggested process is equivalent to the maximum dependency criterion for feature selection, without, however, the need of the multivariate probability densities estimation.

The feature selection approach suggested in the thesis is applied on a number of audio classification tasks, including speech detection in broadcast news and voice pathology detection and discrimination from vowel recordings. The complementarity of the modulation spectral features to the state-of-the-art Mel frequency cepstral coefficients is shown for the above classification tasks. A system for the automatic discrimination of pathological heart murmurs using a high resolution time-frequency analysis of the phonocardiogram (PCG) is also presented. The classification accuracy of the system is comparable to the diagnostic accuracy of experienced paedo-cardiologists on the same PCG dataset.

Περίληψη

Για την περιγραφή των μη στάσιμων ηχητικών σημάτων, έχουν προταθεί διάφορες αναπαραστάσεις συχνότητας - χρόνου και τρόποι υπολογισμού των αντίστοιχων φασμάτων διαμόρφωσης πλάτους (AM) ή συχνότητας (FM). Η πρόοδος που έχει σημειωθεί στις αναπαραστάσεις αυτές, επιτρέπει την καλύτερη απεικόνιση της λεπτής δομής και της δυναμικής της κυματομορφής του σήματος. Ωστόσο η ολοένα μεγαλύτερη ευκρίνεια αυτών των μοντέλων, αυξάνει το πλήθος των παραμέτρων και τη στατιστική τους διακύμανση, και δυσχεραίνει την ταξινόμηση των σημάτων.

Σε αυτή την εργασία μελετάμε το πρόβλημα της μείωσης των διαστάσεων αυτών των αναπαραστάσεων και της επιλογής των 'βέλτιστων' χαρακτηριστικών ως προς την ταξινόμηση που μας ενδιαφέρει. Το κριτήριο είναι η αμοιβαία πληροφορία ως προς τη μεταβλητή της κατηγοριοποίησης. Προτείνουμε την επιλογή των 'βέλτιστων' από τα χαρακτηριστικά που προκύπτουν μετά τη διάσπαση (τον αποκλεισμό) των ιδιόμορφων τιμών υψηλής τάξης της αναπαράστασης. Δείχνουμε ότι αυτή η διαδικασία επιλογής χαρακτηριστικών ικανοποιεί το κριτήριο της μέγιστης εξάρτησης των μεταβλητών από την κλάση ταξινόμησης - με το πλεονέκτημα ότι δε χρειάζεται να υπολογίσουμε κατανομές πιθανότητας πολλών μεταβλητών.

Οι εφαρμογές που περιγράφουμε αφορούν την ανίχνευση ομιλίας σε τηλεοπτικά ή ραδιοφωνικά δελτία ειδήσεων και την ανίχνευση και τη διάκριση διαφόρων παθολογιών της φωνής με βάση τα φάσματα διαμόρφωσης πλάτους (AM). Στις εφαρμογές αυτές δείχνουμε ότι τα επιλεγμένα χαρακτηριστικά μπορούν να βελτιώσουν σημαντικά την απόδοση ενός συστήματος ταξινόμησης σε συνδυασμό με τους ευρέως χρησιμοποιούμενους συντελεστές Mel frequency cepstral coefficients. Τέλος περιγράφουμε ένα σύστημα αυτόματου εντοπισμού των παθολογικών φυσημάτων της καρδιάς, το οποίο βασίζεται σε μια αναπαράσταση συχνότητας - χρόνου πολύ υψηλής ανάλυσης των φωνοκαρδιογραφημάτων. Δείχνουμε ότι η ακρίβεια του συστήματος είναι συγκρίσιμη με τη διαγνωστική ακρίβεια έμπειρων παιδοκαρδιολόγων στην ταξινόμηση των ίδιων φωνοκαρδιογραφημάτων.

Ευχαριστίες

Η διατριβή αυτή είναι το αποτέλεσμα της εργασιακής εμπειρίας 5 χρόνων σαν υποφήφια διδάκτορας στο Εργαστήριο Πολυμέσων του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης (Π.Κ.). Υποστηρίχθηκε οικονομικά από τον Ειδικό Λογαριασμό Κονδυλίων Έρευνας (Ε.Λ.Κ.Ε.) του Π.Κ. και το Ινστιτούτο Πληροφορικής του Ι.Τ.Ε.

Θέλω να ευχαριστήσω ιδιαίτερα τον επόπτη μου, καθηγητή στο Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης, κ. Γ. Στυλιανού για την επιστημονική καθοδήγηση, την ηθική συμπαράσταση και την υπομονή του μαζί μου, σε όλη τη διάρκεια αυτής της εργασίας. Ήταν τιμή μου να δουλεύω μαζί του. Ευχαριστώ επίσης τα μέλη της τριμελούς επιτροπής αυτής της διδακτορικής διατριβής, κ. Πέτρο Μαραγκό, καθηγητή στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, και κ. Γιώργο Τζιρίτα, καθηγητή στο Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης, για τις εύστοχες και εποικοδομητικές παρατηρήσεις τους στις πρώτες παρουσιάσεις αυτής της εργασίας.

Το περιβάλλον εργασίας στο κτίριο του Πανεπιστημίου Κρήτης στην Κνωσσό, ήταν ιδιαίτερα ευχάριστο και δημιουργικό χάρη στον κ. Στυλιανού και στα άλλα μέλη του Εργαστηρίου Πολυμέσων: τους μεταδιδακτορικούς πλέον ερευνητές Γιάννη Πανταζή, Andre Holzapfel, και Γιάννη Αγιομυργιαννάκη, τους πρώην μεταπτυχιακούς φοιτητές Μίλτο Βασιλάκη, Μαρία Αστρινάκη, Χριστίνα Λιονουδάκη, Γιώργο Καφετζή, Μαρία Κουτσογιαννάκη, Γιώργο Γκρέκα, Γιώργο Τζεδάκη, Χρήστο Τζαγκαράκη (κάποιοι συνεχίζουν ως διδακτορικοί φοιτητές πλέον), καθώς και τους φοιτητές από άλλα ευρωπαικά Πανεπιστήμια που δούλεψαν για κάποιους μήνες στο εργαστήριό μας: τον Michael Wohmayr, τον Michael Stark, τον Julian Arias-Londono, τον Robert Peharz. Είχα την τύχη επίσης να έχω καθημερινή παρέα στα διαλείμματα τη Roza Akkus και την Ευγενία Ταμπακάκη - και οι δυο δούλευαν σε διπλανά γραφεία και είχαμε συγχρονίσει τις ώρες εργασίας και καφέ.

Τέλος, ευχαριστώ την οικογένειά μου, Γιώργο, Μανόλη και Κωστή Δαμασκηνάκη, για την υποστήριξη και τη χαρά που μου δίνουν πάντα.

Contents

List of Figures

List of Tables

Acronyms

AM	Amplitude Modulation
CPU	Central Processing Unit
DET	Detection Error Trade-off
DFT	Discrete Fourier Transform
DSTFT	Discrete Short Time Fourier Transform
ECG	ElectroCardioGram
EER	Equal Error Rate
FFT	Fast Fourier Transform
FM	Frequency Modulation
GMM	Gaussian Mixture Model
GRBAS	Grade, Roughness, Breathiness, Asthenicity, Strain
HOSVD	Higher Order Singular Value Decomposition
IB	Information Bottleneck
MFCC	Mel Frequency Cepstral Coefficients
MI	Mutual Information
mRMR	minimum Redundancy Maximum Relevance
PCG	PhonoCardioGram
ROC	Receiver Operating Characteristic
SNR	Signal-to-Noise Ratio
SV	Singular Vector
SVD	Singular Value Decomposition
SVM	Support Vector Machine

Chapter 1

Introduction

The auditory system of humans and animals, can efficiently extract the behaviorally relevant information embedded in natural acoustic environments. Evolutionary adaptation of the neural computations and representations, has probably facilitated the detection of such signals with low SNR over natural, coherently fluctuating background noises [12, 97]. It has been argued [121] that the statistical analysis of natural sounds - vocalizations, in particular - could reveal the neural basis of acoustical perception. Insights in the auditory processing then, could be exploited in engineering applications for efficient sound identification, e.g. speech discrimination from music, animal vocalizations and environmental noises.

Similar to a Fourier analyser, our auditory system maps the one-dimensional sound waveform to a two-dimensional time-frequency representation through the cochlea, the hearing organ of the inner ear [135]. In mammals cochlea is a fluid filled coiled tube, about one cubic centimetre in volume, which resembles the shell of a sea-snail. The eardrum transmits the incoming sounds to the cochlear fluids as pressure oscillations. These oscillations in turn, deflect a membrane of graded mechanical properties which runs along the cochlear spiral, the basilar membrane (BM). Stiffness grading of the basilar membrane serves to analyze incoming sounds into their component frequencies from 20 to 20,000 Hz. A different resonant-like frequency response characterizes each place along the membrane: peak frequencies and bandwidths are highest near the spiral's broad mouth (where the BM is stiffest), and lowest further up the tube near the spiral's apex [135, 107].

The tube in addition carries neurosensory hair cells that fire in response to vibrations in the thin cochlear fluid; nerve impulses are then sent to the brain as electrical signals. Manoussaki et al [76, 75] have found that the energy of the sound waves travelling in the cochlea is not evenly distributed along the tube; the spiral shape of the cochlea focuses wave energy towards the outer

wall, especially further up the tube, where lower frequencies (bass sounds) are detected [76, 75]. This concentration of energy in the spiral's outer edge, amplifies the sensitivity of membrane cells to vibrations. This amplification corresponds to a 20 decibels boosting of lower frequencies relative to the higher frequencies detected at the outer face of the spiral. Mammals may have developed the spiral structure for communication and survival, since low frequency sound waves can travel further [76, 75].

Recent findings from auditory physiology, psychoacoustics, and speech perception, suggest that the auditory system re-encodes acoustic spectrum in terms of spectral and temporal modulations. The perceptual role of very low frequency modulations resembles inaudible "message bearing waves" modulating higher frequency carriers. The existence of modulations in speech has been evidenced in [77, 102]. Dynamic information provided by the modulation spectrum includes fast and slower time-varying quantities such as pitch, phonetic and syllabic rates of speech, tempo of music, etc [8]. Moreover, speech intelligibility depends on the integrity of the slow spectro-temporal energy modulations.

Accordingly, the auditory models proposed by various researchers, are based on a two stage processing: the early stage consists of spectrum estimation whereas in the second "cortical" stage, spectrum analysis is performed in order to estimate low frequency modulations. Various models of modulation representations have been proposed: a two-dimensional Fourier transform of auto-correlation matrix of sound spectrogram [121], a two-dimensional wavelet transform of auditory spectrogram [137], a joint acoustic / modulation frequency representation [8], a combination of the latter representation with cepstrum [8], etc. In particular the two latter representations have been explored in this thesis (cf Chapter 2).

Auditory models are perhaps too complex for successful inversion which would be required for coding, speech enhancement, compression, etc. For recognition and classification systems however, robustness in noise and utility for segregation in cluttered acoustic environments might be more important than invertibility.

In order for these features to describe natural signals like speech, we might have to consider fine resolution in all the subspaces involved. Using more relevant features from the original space is not satisfactory due to the fact that computation cost becomes very high and the generalization of classification algorithm is not guaranteed: the curse of dimensionality [30, 15] hinders the classification task. There is a trade-off between compactness and quality of a representation - measured in terms of redundancy and relevance of features, respectively. The relevance-redundancy trade-off can be explored using an information theoretic concept, mutual information [22]. In

the supervised learning framework, features are regarded as relevant if they provide information about a target. Assuming that this additional variable Y (the target) is available, relevance of features to a given discrimination task can be assessed based on their mutual information to the target class [22]. Redundancy on the other hand can be also defined based on the mutual information between features.

Both redundancy and relevance of features need to be addressed before selecting points in the feature space for an efficient, robust automatic classification system. Depending on the task, data available for training and features extracted, a different trade-off occurs. Redundant (as opposed to compact) representations might be an advantage in the presence of noise and uncertainty [13]. Different feature selection methods have been proposed in the past in order to extract the time-varying information of interest to the audio classification task. Two of them are presented in Chapter 3: information bottleneck method [128] and maximum relevance-minimum redundancy feature selection scheme [101]. In addition, we propose a different approach which considers both redundancy reduction and relevance maximization; redundancy is reduced first, before searching for relevant features, using a technique from multilinear algebra, the higher order singular value decomposition (HOSVD [68], cf Chapter 3). We experimentally assess the usefulness of this approach in different audio classification tasks in subsequent chapters through cross-validation procedures.

Contributions and structure of the Thesis

The main contribution of this thesis is related to the feature selection for classification of the (multidimensional) representations of audio signals. Specifically, we show that the maximal dependency feature selection can be approximated by a maximum relevance feature selection scheme applied to the least redundant features: minimum redundancy is achieved through higher order SVD whereas the selection of the most relevant features proceeds next based on mutual information estimation. Model selection is performed through cross-validation procedure using SVM classifier. The most successful applications of this technique include the use of modulation spectra for speech detection in Chapter 4, and for voice pathology detection and classification in Chapter 5. We also show their complementarity to the state-of-the-art MFCC features for the same task. In Chapter 6 we extract features from a high resolution time-frequency representation of digital recordings of heart sounds (phonocardiograms, PCG); based on these features, heart murmur classification accuracy is shown to be comparable to the diagnostic accuracy of

experienced paedo-cardiologists on the same PCG dataset.

The structure of this thesis is divided into the following chapters:

- Two-dimensional energy representations of signals constitute the basic method to analyze and describe non-stationary audio signals whose energy distribution varies over frequency (or scale) and time, such as speech, music, etc. Chapter 2 briefly reviews the most common energy representations of audio signals and the techniques used for subsequently analyzing these representations and detecting amplitude (or frequency) modulations.

- Chapter 3 addresses the issue of feature selection from multi-dimensional signal representations. We briefly review the information bottleneck (IB) method and we describe its application on the auditory model of Shamma et al [137] towards speech discrimination. The maximum relevance-minimum redundancy feature selection scheme [101] is described next. We propose a different approach which first considers redundancy reduction, using a technique from multilinear algebra, the higher order singular value decomposition (HOSVD). We proceed then to select the most relevant of features based on their mutual information to the class variable. Next chapters present applications of this approach on modulation spectra and reassigned spectrogram - a high resolution time-frequency representation - for various classification tasks. Each of these "technical" chapters is preceded by a brief synopsis of its contents.

- Automatic detection of speech finds applications in human-machine interfaces, multimedia processing for automatic labelling and extraction of semantic information. Chapter 4 presents experiments on speech discrimination in broadcast news from Greek and US English tv and radio programs, based on modulation spectra and their combination with MFCC. We also mention additional experiments we conducted on singing voice detection as well as on the evaluation of modulation spectra for speaker identification or verification.

- Many studies have focused on identifying acoustic measures that highly correlate with pathological voice qualities (also referred to as voice alterations). Using acoustic analysis, the degree of voice alterations can be objectively evaluated in a noninvasive manner. Chapter 5 explores the information provided by the joint acoustic and modulation frequency representation for the detection and discrimination of voice disorders. A normalization technique is proposed for the modulation spectral features which we use in cross-database experiments.

- The invention of heart signal analysis can be traced back to Hooke who first suggested in the 1700s that it might be possible to detect various malfunctions by the analysis of the sound they make. Classic heart auscultation using a conventional stethoscope to detect abnormal heart sounds is the most common and widely recommended method nowadays to screen for structural abnormalities of the cardiovascular system [57, 92]. Detecting relevant symptoms and forming a diagnosis based on the sounds heard through a stethoscope, however, is a skill that can take years to acquire and refine [44]. Chapter 6 describes a system that automatically discriminates between normal and pathological systolic heart murmurs in children. The features are extracted from a detailed time-frequency representation of the phonocardiogram, the digital recording of the heart sounds obtained by an electronic stethoscope.

- Finally Chapter 7 concludes and summarizes the major contributions of this thesis.

Chapter 2

Feature Extraction from Audio Signals

2.1 Non-stationary Signal Analysis

One of the most important aspects in signal processing is the extraction of useful information from signals by transforming them. If the transformation is invertible, it unambiguously represents the signal; operations such as pattern recognition for classification, can be performed then on the transformed signal, where the relevant characteristics might be more clear [114].

2.1.1 Time-frequency distributions

The Fourier transform is the most common invertible transform for *stationary* signals, whose properties remain constant in time. For such signals $x(t)$, Fourier transform (FT) is defined as:

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp^{-2j\pi ft}\, dt \tag{2.1}$$

where the analysis coefficients $X(f)$ are the inner products of the signal with sinewave basis functions of *infinite* duration. Therefore, transforming a non-stationary signal through FT, would extend any change in time over all frequencies f in $X(f)$ ([114] and references therein).

One possible adaptation of Fourier transform (FT) for the analysis of a non-stationary signal $x(t)$, is a "local" FT looking at the signal through a short window $g(t)$ over which the signal is almost stationary. Gabor first adapted, in the 1940s, the FT of the windowed signal $x(t)g^*(t-\tau)$,

centered at time instant τ, to yield the Short-Time Fourier Transform (STFT) [114]:

$$\mathrm{STFT}(\tau, f) = \int x(t) g^*(t-\tau) \exp^{-2j\pi f t}\, dt. \tag{2.2}$$

STFT maps the signal $x(t)$ into a *two-dimensional* time-frequency representation consisting of the time-dependent spectral characteristics of the signal. The squared magnitude of STFT is the *spectrogram*. Actually, Gabor defined in 1946 the synthesis formula which corresponds to the analysis formula in eq. 2.2 [114]. The analysis of human speech was the main reason for the practical development in the 1940s of time-frequency analysis [107]); the main method was and still is the STFT.

The choice of the window $g(t)$ critically affects the properties of the STFT analysis regarding its resolution in time and frequency. According to eq. 2.2, there is a dual interpretation of STFT. At first, STFT can be viewed as a succession of FTs: for a windowed segment of the signal around a given time t, every "frequency" of the STFT is estimated. Or, given a frequency f, the whole signal is filtered with a bandpass filter whose impulse response is the window $g(t)$ modulated to that frequency. STFT can be considered then a modulated analysis filterbank [114].

Given the window function $g(t)$ and its FT $G(f)$, two pure sinusoids with frequencies f_1, f_2 can be discriminated only if they differ more than Δf (STFT resolution in frequency):

$$\Delta f^2 = \frac{\int f^2 |G(f)|^2 df}{\int |G(f)|^2 df} \tag{2.3}$$

where Δf is the "bandwidth" of the filter (in the modulated filterbank) and the denominator is the energy of the signal. In addition, two acoustic events at times t_1, t_2 can be discriminated only if $|t_1 - t_2| \geq \Delta t$ (STFT resolution in time):

$$\Delta t^2 = \frac{\int t^2 |g(t)|^2 dt}{\int |g(t)|^2 dt} \tag{2.4}$$

where Δt is the spread in time (related to the length of the analysis window $g(t)$) and the denominator is the energy of the signal.

The time-bandwidth product is lower bounded as:

$$\Delta t \Delta f \geq \frac{1}{4\pi} \tag{2.5}$$

with equality satisfied when Gaussian windows are used. This is the "acoustic uncertainty prin-

ciple" [61] or Heisenberg inequality which states that the more precise the measurement of an acoustic event in time, the less precise its simultaneous measurement in frequency domain is; and vice-versa [114]. When $g(t)$ is short (wideband analysis) the frequency subbands will be wide and the time resolution will be high. When $g(t)$ is long (narrowband analysis) the frequency subbands will be narrow and the time resolution will be low. Once a window has been chosen, STFT represents the spectral characteristics of the windowed signal as a set of point estimates on the time-frequency lattice, with *fixed* coordinates determined by the length and overlap of the analysis window. The time and frequency resolution of STFT then is *constant* over the time-frequency plane [114]. Alternatively, one could let the time and frequency resolution vary in the time-frequency plane for a multi-resolution analysis [114].

2.1.2 Time-scale distributions

Speech and music exploit different ranges of temporal and spectral processing. The auditory cortices have been accordingly differentiated [144]. In order to adapt to the "acoustic uncertainty principle", right and left auditory cortices exhibit differences between spectral and temporal resolution, presumably due to different temporal integration windows ([144] and references within). These conclusions were supported by magnetoencephalography (MEG) data and electrical potential recordings from the human auditory cortex [61]. The left auditory cortex has a short analysis window (20-50 ms) which is optimal for detecting fine temporal differences (relevant for speech distinctions), at the expense of broader frequency tuning. The analysis window in the right cortex is longer (150-250 ms) which enables a greater spectral resolution (sharper frequency tuning) but limits the sensitivity to rapid acoustic cues.

The one-dimensional sound waveform is first mapped to a two-dimensional time-frequency representation through the cochlea, the hearing organ of the inner ear [135]. A membrane of graded mechanical properties runs along the cochlear spiral, the basilar membrane (BM). Input stimuli cause the basilar membrane to vibrate; due to the grading of membrane stiffness, incoming sounds are analyzed into their component frequencies from 20 to 20,000 Hz. Moreover, a different resonant-like frequency response characterizes each place along the membrane: peak frequencies and bandwidths are highest near the spiral's broad mouth (where the BM is stiffest), and lowest further up the tube near the spiral's apex [135, 107]. These auditory filters achieve approximately constant-Q resolution, i.e., the resolution in frequency Δf decreases whereas resolution in time Δt increases with the central frequency f of the auditory band-pass filters [107]. We can therefore

define Δf as proportional to f:

$$\frac{\Delta f}{f} = c \tag{2.6}$$

where c is a constant [114]. The frequency responses of the analysis filters are regularly spread in a *logarithmic scale* along the frequency axis, instead of the uniform coverage for the STFT analysis.

When the relative bandwidth $\Delta f/f$ of the analysis filters is constant, Δf and Δt still satisfy the Heisenberg inequality (eq. 2.5) but the time resolution can be very good at high frequencies whereas frequency resolution can be very good at low frequencies. This analysis is most suitable then for signals consisting of short, high frequency components and long-lasting, low frequency components [114].

The front-end auditory processing along the basilar membrane can be modeled as a *Continuous Wavelet Transform* (CWT): the impulse responses of the filter bank are defined as *scaled* (expanded or contracted) and translated versions of the same prototype, the *basic wavelet* $h(t)$ [107, 114]:

$$h_\alpha(t) = \frac{1}{\sqrt{|\alpha|}} h\left(\frac{t}{\alpha}\right) \tag{2.7}$$

where α is the scale factor; CWT is defined then as [114]:

$$\mathrm{CWT}_x(\tau, \alpha) = \frac{1}{\sqrt{|\alpha|}} \int x(t) h^*\left(\frac{t-\tau}{\alpha}\right) dt \tag{2.8}$$

The *scalogram*, or *wavelet spectrogram*, is defined as the squared modulus of the CWT. It is the distribution of the energy of the signal in the time-scale plane, with different resolutions in contrast to the spectrogram [114]. We need phase information to reconstruct the signal from both energy representations; also, these representations are bilinear functions of the analysed signal and contain thus cross-terms due to interference phenomena [114]. More advanced time-frequency and time-scale representations have been developed and a link between spectrogram, scalogram and Wigner-Ville distribution has been established, merging them into a common class of energy representations (see [67, 114] and references therein).

2.1.3 Reassignment of time-frequency and time-scale representations

Reassignment aims at sharpening time-frequency and time-scale representations in order to improve their readability (see [9, 40, 41] for more details on the reassignment method). A spectrogram contains just a few cross-terms arising from interference phenomena; however, it is

constrained by the trade-off between resolution in time and frequency, which results in poor concentration and resolution of the individual components of multi-component signals [41]. The reassignment method was specifically designed to overcome this limitation.

Let a short-time Fourier transform (STFT) of a signal $x(t)$, denoted as $\mathcal{F}_x(t,\nu;g)$ where t, ν refer to time and frequency, respectively, and g is the analysis window. A reassigned version of the spectrogram is obtained by shifting each value $S_x(t,\nu;g)$ of the spectrogram (spectral energy - squared modulus of the STFT) computed at a point (t,ν) of the time-frequency lattice, to another point $(\hat{t},\hat{\nu})$ which is the center of gravity of the energy distribution around (t,ν). The value of the reassigned spectrogram $S_x^{(r)}(t',\nu';g)$ at a point (t',ν') is the sum of all the spectrogram values *reassigned* to this point:

$$S_x^{(r)}(t',\nu';g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} S_x(t,\nu;g)\delta(t'-\hat{t}(x;t,\nu))\delta(\nu'-\hat{\nu}(x;t,\nu))dtd\nu \tag{2.9}$$

The reassigned time-frequency distribution uses information from the phase spectrum to sharpen the amplitude estimates for each bin of the STFT [51], since the operators $\hat{t}(x;t,\nu)$ and $\hat{\nu}(x;t,\nu)$ can be expressed as:

$$\begin{aligned} \hat{t}(x;t,\nu) &= -\frac{d\Phi_x(t,\nu;g)}{d\nu} \\ \hat{\nu}(x;t,\nu) &= \nu + \frac{d\Phi_x(t,\nu;g)}{dt}, \end{aligned} \tag{2.10}$$

where $\Phi_x(t,\nu;g)$ is the phase of the STFT of x:

$$\Phi_x(t,\nu;g) = \arg(\mathcal{F}_x(t,\nu;g)). \tag{2.11}$$

In practice, somewhat different expressions are used for the reassignment operators leading to an efficient implementation (see the Tutorial and Reference guides in [10]).

2.2 Spectrum Analysis

Speech is characterized by the joint spectro-temporal energy modulations in its spectrogram; these oscillations in power across spectral and temporal axes, reflect the formant peaks, their transitions, spectral edges, and fast amplitude modulations at onsets-offsets [107]. Evidence for the existence of speech modulations has been provided in [77, 102]. Of particular relevance to speech intelligibility, are the slowly varying amplitude and frequency modulations of sound [143].

Slow temporal modulations (few Hz) correspond to the phonetic and syllabic rates of speech [107]. Measurement of detection thresholds for these spectro-temporal modulations have revealed the lowpass character - in both dimensions - of the modulation transfer functions (MTF) of our auditory system [20]: 50% bandwidths of 2 cycles/octave and 16 Hz over the perceptually important range of $0.25 - 8$ cycles/octave and $1 - 128$ Hz, respectively.

Various models of modulation representations have been proposed in the literature in order to simulate sound processing in the auditory system: a two-dimensional Fourier transform of autocorrelation matrix of the sound spectrogram [121], a two-dimensional wavelet transform of auditory spectrogram [137], a joint acoustic / modulation frequency representation [8], a combination of the latter representation with cepstrum [8], etc. The two latter representations are briefly presented below. For completeness, we also mention the Energy Separation algorithm based on Teager energy operator.

2.2.1 Computational Auditory Model

Shamma et al [141, 137] have proposed a computational auditory model based on wavelet decomposition [114], which reproduces the main trends in the experimentally determined spectro-temporal modulation transfer functions (MTFs) [20]. Early stages of the model consist of inner ear processes (in the cochlea and the auditory nerves up to the cochlear nucleus) in order to estimate an enhanced spectrogram of incoming sounds. During later stages spectrum analysis occurs: fast and slow modulation patterns are detected by arrays of filters centered at different frequencies, with Spectro-Temporal Response Functions (STRFs) resembling the receptive fields of auditory midbrain neurons [34, 105]. These filters have the form of a spectro-temporal Gabor function, selective for specific frequency sweeps, bandwidth, etc., performing actually a multi-resolution wavelet analysis of the spectrogram in order to provide a suitable basis set for speech stimuli [137]. The auditory based features are collected from an audio signal in a frame-per-frame scheme. For each time frame, the auditory representation is calculated on a range of frequencies, scales (of spectral resolution) and rates (temporal resolution). The extracted information is averaged over a fixed time window (e.g., of 500ms), therefore resulting in a 3-dimensional array, or third-order tensor. The model has been successfully applied in the assessment of speech intelligibility [34], the discrimination of speech from non-speech [95, 138, 91], as well as in simulations of various psychoacoustical phenomena [18].

Joint acoustic / modulation frequency representations and their combination with cepstrum [8], represent a simpler interpretation of the auditory model and it is presented in the following sec-

tion.

2.2.2 Modulation Frequency Analysis

The most common modulation frequency analysis framework [50, 8] for a discrete signal $x(n)$, initially computes via the discrete Fourier transform (DFT) the discrete short-time Fourier transform (DSTFT) $X_k(m)$, m denoting the frame number and k the DFT frequency sample:

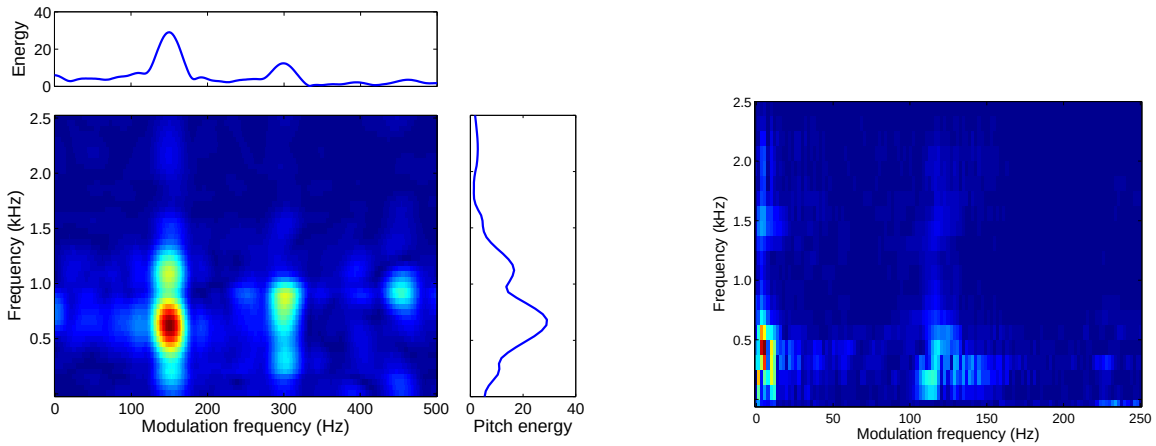
$$\begin{aligned} X_k(m) &= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^{kn}, \\ k &= 0,\ldots,K-1, \end{aligned} \tag{2.12}$$

where $W_K = e^{-j(2\pi/K)}$, $h(n)$ the (acoustic) frequency analysis window and M the hopsize (in number of samples). The mean can be subtracted from each subband envelope - defined as the magnitude $|X_k(m)|$ or square magnitude $|X_k(m)|^2$ of the subband - before modulation frequency estimation, in order to reduce the interference of large DC components (of subband envelopes). Subband envelope detection and their frequency analysis (with DFT) are performed next, to yield the modulation spectrum with a uniform modulation frequency decomposition:

$$\begin{aligned} X_l(k,i) &= \sum_{m=-\infty}^{\infty} g(lL-m)|X_k(m)|W_I^{im}, \\ i &= 0,\ldots,I-1, \end{aligned} \tag{2.13}$$

where $W_I = e^{-j(2\pi/I)}$, $g(m)$ is the modulation frequency analysis window and L the corresponding hopsize (in number of samples); k and i are referred to as the "Fourier" (or acoustic) and "modulation" frequency, respectively. Tapered windows $h(n)$ and $g(m)$ are used to reduce the sidelobes of both frequency estimates.

The magnitude of the acoustic-modulation frequency representation computed in eq. 2.13 is referred to as modulation spectrogram. It displays the modulation spectral energy $|X_l(k,i)| \in R^{I_1 \times I_2}$ in the joint acoustic/modulation frequency plane. Length of the analysis window $h(n)$ controls the trade-off between resolutions in the acoustic and modulation frequency axes [120]. When $h(n)$ is short (wideband analysis) the frequency subbands will be wide and the maximum observable modulation frequency will be high. When $h(n)$ is long (narrowband analysis) the frequency subbands will be narrow and the maximum observable modulation frequency will be low. The degree of overlap between successive windows sets the upper limit of the subband sampling rate during the modulation transform.



(a) (b)

Figure 2.1: (a) Modulation spectrogram of sustained vowel /AH/ by a 34 years old normal male speaker ($\sim$ 150 Hz fundamental frequency). The two side plots present the slices intersecting at the point of maximum energy; its coordinates coincide with the fundamental frequency and the first formant of /AH/ ($\sim$ 590 Hz). Vertical plot displays the localization of fundamental frequency energy at vowel formants along the acoustic frequency axis; the upper horizontal plot presents the energy localization of first formant at the fundamental frequency and its harmonics along the modulation frequency axis. (b) Modulation spectrogram $|X_l(k,i)|$ for 500 ms of a speech signal. Energy at modulations corresponding to pitch ($\sim$ 120 Hz) and syllabic and phonetic rates ($<$ 40 Hz) remain prominent.

Fig. 2.1a shows the modulation spectrogram $|X_l(k,i)|$ of a 262 ms long frame from sustained phonation speech samples of the vowel /AH/ uttered by a normal male speaker from the MEEI database [37] (cf Chapter 5 on voice pathology detection). Apparently this phonation does not possess the syllabic and phonetic temporal structure of speech (compare to Figure 2.1b). Hence, the higher energy values are not concentrated at the lower modulation frequencies which are typical in running speech, $\sim 1-20$ Hz [50]. Instead, since we used an analysis window $h(n)$ that was shorter than the expected lowest pitch period, the highest energy terms usually occur at the fundamental frequency of the speaker ($\sim$ 150 Hz) and its harmonics in the modulation frequency axis (up to 500 Hz). Fundamental frequency energy appears localized at the first two formants of vowel /AH/ along the acoustic frequency axis (their range is $\sim 677 \pm 95$ Hz and $\sim 1083 \pm 118$ Hz). Figure 2.1b depicts the modulation spectrum for 500 ms of a speech signal. Energy at modulations that characterize speech at the lower acoustic frequency bands, corresponding to syllable and phonemic rates ($<$ 40 Hz) and the pitch of speaker ($\sim$ 120 Hz), remain prominent.

2.2.3 The Teager Energy Operator

For completeness, we briefly mention another approach to the estimation of AM-FM modulations, the high resolution Teager energy operator (see [77, 78, 79, 107] and references therein). It has been shown in [78, 79] that the *discrete-time energy operator*, defined as:

$$\Psi_d(x[n]) = x^2[n] - x[n-1]x[n+1], \tag{2.14}$$

when applied to a real-valued AM-FM discrete signal of the form:

$$x[n] = \alpha[n]\cos(\phi[n]) = \alpha[n]\cos\left(\omega_c n + \omega_m \int_0^n q[m]dm + \theta\right) \tag{2.15}$$

with instantaneous frequency:

$$\omega_i[n] = \frac{d}{dn}\phi[n] = \omega_c + \omega_m q[n] \tag{2.16}$$

tracks the energy of the signal, which is equal to the squared product of the amplitude envelope and the instantaneous frequency, i.e., the AM and FM components:

$$\Psi_d(x[n]) \approx \alpha^2[n]\sin^2(\omega_i[n]) \tag{2.17}$$

where $q[n] \leq 1$, $\omega_m \in [0, \omega_c]$ is the maximum deviation from the carrier frequency ω_c, θ is a constant phase onset, and the signals $\alpha[n]$ and $\omega_i[n]$ do not vary too fast or too much in time compared to ω_c [77, 78, 79, 107] .

The Teager energy operator was formulated in the context of an harmonic oscillator formulation, where a speech resonance (formant) is referred to as an oscillator system - formed by the vocal tract cavities - which emphasizes or de-emphasizes certain frequencies even within a single pitch period (see [107] and references therein). Teager energy operator is almost instantaneous and captures energy fluctuations within a glottal cycle, provided that the resonant components have been separated before, through bandpass-filtering around the center frequency of each resonance [107]. The amplitude envelope $\alpha[n]$ and the instantaneous frequency $\omega_i[n]$ in eq. 2.17 can be interpreted then as the amplitude and frequency modulations of the cavity resonances (speech formants modelled as AM-FM sinewaves).

It is possible then to separate these AM and FM components through continuous or discrete-time *energy separation algorithms* based on the Teager energy operator [77, 107].

Chapter 3

Dimensionality Reduction and Feature Selection for Classification

Section 3.3 of the chapter has included parts of the following publications:

- Markaki M., Wohlmayr M. and Stylianou Y., *Speech - Nonspeech Discrimination using the Information Bottleneck Method and Spectro-Temporal Modulation Index*, InterSpeech ICSLP, 2007.
- Wohlmayr M., Markaki M., and Stylianou Y., *Speech - Nonspeech Discrimination based on Speech-relevant Spectrogram Modulations*, EUSIPCO, 2007.
- Markaki M., Wohlmayr M. and Stylianou Y., *Extraction of Speech-Relevant Information from Modulation Spectrograms*, Progress in Nonlinear Speech Processing, Springer Berlin / Heidelberg, Springer, pp. 78 - 88, 2007.

Many thanks to Michael Wohlmayr (Graz University of Technology) for his collaboration during his stay in University of Crete as an Erasmus MSc student.

3.1 Introduction

Classification algorithms detect and exploit complex patterns in data during training, validation and testing. Curse of dimensionality refers to the phenomenon that a large class of learning algorithms can fail when applied on high-dimensional problems due to the following reasons [30] (www.support-vector.net):

- The representation of complex patterns poses a significant computational problem (in terms of computational cost and storage volumes) when working with very large vectors.
- The exclusion of accidental, unstable patterns which might lead to overfitting, is a difficult statistical problem in high dimensional feature spaces. False regularities in the training set due to noise, for example, would not be repeatedly found in a test set. The over-fitting of the training system, results in the generalization error increasing with the effective dimension N of the data; and the number of training examples required for achieving a given error level, can grow exponentially with the number of dimensions [30, 15].

In order to improve generalization, we need a more useful low-dimensional representation of the data. To this end, both dimensionality reduction and feature selection techniques can be employed; these techniques differ in the selection of unsupervised or supervised learning algorithms, respectively. Unsupervised dimensionality reduction can preserve information from all the original input variables, promoting generalization. On the other hand, a purely unsupervised technique might throw away low variance dimensions that are highly predictive for a given classification task.

Representation of an audio signal might be a function of two variables like time-frequency distributions [67] and modulation spectra [8], or even three variables as in multi-scale spectro-temporal modulations [137]. For the dimensionality reduction of sound representations we can employ a generalization of Singular Value Decomposition (SVD) algorithm to tensors (multi-dimensional matrices), referred to as Higher Order SVD [68]. The singular value decomposition of a time-frequency distribution was first proposed in [80] for the Wigner distribution. HOSVD of multiscale spectro-temporal modulations was first done in [96].

HOSVD enables the decomposition of tensors to their singular vectors [68]. There are three reasons for performing such a decomposition:

- Real signals contain noise; noise is typically spread out over all the terms of the HOSVD decomposition whereas signals are well represented by the first few terms. Truncating

the series after the first few terms, significantly reduces noise while retaining most of the signal [67].

- The signal representations can be approximated in a lower-dimensional space producing a compact feature set suitable for classification. The term "compact" also refers to the fact that the HOSVD addresses features redundancy by selecting mutually independent features [68].

- Comparing to SVD, reduction in dimensionality of feature space through HOSVD can be performed in every subspace separately.

However, through HOSVD method, the low variance projections are discarded without testing if these could be useful for classification.

In the supervised learning framework, on the other hand, features are regarded as relevant if they provide information about a target. In the case of speech processing systems, the available tagging y of the audio signal (as speech / non speech class, normal / pathological voice, etc.) guides the selection of features during training. The relevance of information in the representation of an audio signal is given by Shannon's mutual information [22]. We tested information bottleneck (IB) method [128] towards the classification of Multi-Scale Spectro-temporal Modulations [137] for the speech detection task. Through IB method, we defined the relevant modulation spectrum of each sound ensemble, speech and non-speech (music, animal sounds and various noises). Knowledge of the compact modulation patterns allowed us to classify new incoming sounds based on the similarity of their representation to the class-typical patterns. For the similarity assessment, we adapted the spectro-temporal modulation index (STMI) - which is used to the assessment of speech intelligibility [34] - to handle the contribution of different frequency bands. Further, we examined maximum relevance-minimum redundancy (mRMR) feature selection scheme [101]; mRMR considers both redundancy reduction and relevance maximization of features for a given classification task [101]. We proposed a different approach which first reduces features redundancy, using higher order singular value decomposition (HOSVD) [68]. The selection of the most relevant among the least redundant features can proceed then using traditional model selection methods such as cross-validation.

The structure of this chapter is the following. Section 3.2 describes the unsupervised dimensionality reduction technique HOSVD [68]. Section 3.3 presents the information bottleneck (IB) method of Tishby et al [128]. Section 3.4 presents the maximum relevance-minimum redundancy feature selection scheme [101] and our approach. We conclude with a theoretical analysis -

based on the analysis of mRMR scheme in [101] - which shows that the proposed method better approximates the maximum dependency criterion for feature selection [101].

3.2 Higher Order Singular Value Decomposition

We describe the HOSVD implementation process for two-dimensional sound representations like time-frequency distributions and modulation spectra, as reviewed in [68]. Every signal segment in the training database is represented as a two-dimensional matrix. Let I_3 denote the number of signal segments contained in the training set. Thus, I_3 can be seen as a dimension of time (I_1 and I_2 correspond to the acoustic and modulation frequency dimensions, or frequency and time dimensions, respectively). The mean value is then computed over I_3, and it is subtracted from all the spectra in the training set. The zero-mean spectra are then stacked, creating the data tensor $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$.

Higher Order SVD (HOSVD) [68] decomposes a tensor $\mathcal{D}$ to its mode$-n$ singular vectors:

$$\mathcal{D} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \tag{3.1}$$

where $\mathcal{S}$ is the core tensor with the same dimensions as $\mathcal{D}$; $\mathcal{S} \times_n \mathbf{U}^{(n)}$, $n = 1, 2, 3$, denotes the $n-$mode product of $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ by matrix $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$. For $n = 2$ for example, $\mathcal{S} \times_2 \mathbf{U}^{(2)}$ is an $(I_1 \times I_2 \times I_3)$ tensor given by

$$\left(\mathcal{S} \times_2 \mathbf{U}^{(2)}\right)_{i_1 i_2 i_3} \stackrel{\text{def}}{=} \sum_{i_2} s_{i_1 i_2 i_3} u_{i_2 i_2}. \tag{3.2}$$

$\mathbf{U}^{(1)} \in \mathbb{R}^{I_1 \times I_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{I_2 \times I_2}$ are the unitary matrices of the corresponding subspaces of acoustic and modulation frequencies, or frequency and time subspaces respectively; $\mathbf{U}^{(3)} \in \mathbb{R}^{I_3 \times I_3}$ is the samples subspace matrix. These $(I_n \times I_n)$ matrices $\mathbf{U}^{(n)}$, $n = 1, 2, 3$, contain the n-mode singular vectors (SVs):

$$\mathbf{U}^{(n)} = \left[U_1^{(n)}\ U_2^{(n)}\ \ldots\ U_{I_n}^{(n)}\right]. \tag{3.3}$$

Each matrix $\mathbf{U}^{(n)}$ can directly be obtained as the matrix of left singular vectors of the "matrix unfolding" $\mathbf{D}_{(n)}$ of $\mathcal{D}$ along the corresponding mode [68]. Tensor $\mathcal{D}$ can be unfolded to the $I_1 \times I_2 I_3$ matrix $\mathbf{D}_{(1)}$, the $I_2 \times I_3 I_1$ matrix $\mathbf{D}_{(2)}$, or the $I_3 \times I_1 I_2$ matrix $\mathbf{D}_{(3)}$. The n-mode singular values correspond to the singular values found by the SVD of $\mathbf{D}_{(n)}$.

The subtensors $\mathcal{S}_{i-n=\alpha}$ of the core tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, obtained by fixing the nth index

to α, have the properties of [68]:

all orthogonality: two subtensors $\mathcal{S}_{i-n=\alpha}$ and $\mathcal{S}_{i-n=\beta}$ are orthogonal for all possible values of $n, /, \alpha$ and β subject to $\alpha \neq \beta$:

$$\langle \mathcal{S}_{i-n=\alpha}, \mathcal{S}_{i-n=\beta} \rangle = 0 \qquad \text{when} \alpha \neq \beta$$

ordering:

$$\|\mathcal{D}_{i-n=1}\| \geq \|\mathcal{S}_{i-n=2}\| \geq \ldots \|\mathcal{S}_{i-n=2}\| \geq 0$$

for all possible values of n. The Frobenious norms $\|\mathcal{S}_{i-n=i}\|$, symbolized by $\sigma_i^{(n)}$, are $n-$mode singular values of $\mathcal{D}$.

The contribution $\alpha_{n,j}$ of the j^{th} n-mode singular vector $U_j^{(n)}$ is defined as a function of its singular value $\lambda_{n,j}$:

$$\alpha_{n,j} = \lambda_{n,j} \,/\, \sum_{j=1}^{I_n} \lambda_{n,j} \quad \text{or} \quad \alpha_{n,j} = \lambda_{n,j} \,/\, \sqrt{\sum_{j=1}^{I_n} \lambda_{n,j}^2} \tag{3.4}$$

The appropriate values of R_1, R_2 can be determined by inspection of the singular value spectra in the respective modes. By setting a threshold in the contribution of each singular vector, the R_n with $n = 1, 2$ singular vectors (SVs) can be retained for which the contribution exceeds that threshold. Thus, the truncated matrices $\hat{\mathbf{U}}^{(1)} \in \mathbb{R}^{I_1 \times R_1}$ and $\hat{\mathbf{U}}^{(2)} \in \mathbb{R}^{I_2 \times R_2}$ are obtained. Joint acoustic and modulation frequencies or time-frequency representations $\mathbf{B} \in \mathbb{R}^{I_1 \times I_2}$ extracted from audio signals are projected on $\hat{\mathbf{U}}^{(1)}$ and $\hat{\mathbf{U}}^{(2)}$ [68]:

$$\mathbf{Z} = \mathbf{B} \times_1 \hat{\mathbf{U}}^{(1)T} \times_2 \hat{\mathbf{U}}^{(2)T} = \hat{\mathbf{U}}^{(1)T}.\mathbf{B}.\hat{\mathbf{U}}^{(2)} \tag{3.5}$$

where $\mathbf{Z}$ is an $(R_1 \times R_2)-$matrix, and R_1, R_2 denote the number of retained SVs in the acoustic and modulation frequency subspace, or frequency and time subspace, respectively. We can project $\mathbf{Z}$ back into the full $I_1 \times I_2$-dimensional space to get the rank-(R_1, R_2) approximation of $\mathbf{B}$ [68]:

$$\hat{\mathbf{B}} = \mathbf{Z} \times_1 \hat{\mathbf{U}}^{(1)} \times_2 \hat{\mathbf{U}}^{(2)} = \hat{\mathbf{U}}^{(1)}.\mathbf{Z}.\hat{\mathbf{U}}^{(2)T} \tag{3.6}$$

The ordering of the singular values in every mode implies that the "energy" of the tensor $\mathbf{B}$ is mainly concentrated in the part corresponding to low values of indices, i.e., the "significant"

singular values. Consequently, if $\sigma_{R_1}^{(1)} >> \sigma_{R_1+1}^{(1)}$ and $\sigma_{R_2}^{(2)} >> \sigma_{R_2+1}^{(1)}$, the rank-$(R_1, R_2)$ approximation $\hat{\mathbf{B}}$ is a good approximation of the signal with an error bounded by:

$$||\mathbf{B} - \hat{\mathbf{B}}||^2 \leq \sum_{i_1=R_1+1}^{I_1} \sigma_{i_1}^{(1)^2} + \sum_{i_2=R_2+1}^{I_2} \sigma_{i_2}^{(2)^2} \tag{3.7}$$

De Lathauwer and Vandewalle [68] refer that the truncation of the core tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, may destroy its all-orthogonality. If we are interested in best approximation of signals, this approach might be suboptimal comparing to the best rank$-(R_1, R_2)$ approximation described in [68]. For classification however, the selection of the dimensionality of the embedding is a question of model selection (www.support-vector.net). It can be addressed using traditional model selection methods such as cross-validation. According to the "maximum contribution" (or, equivalently, minimum redundancy [101]) criterion, the number of retained components (or SVs) in each subspace can be determined by analyzing the "discriminative" contribution of each component. By including only the components whose contribution is larger than a threshold, the cross-validation classification error (equal error rate, EER, usually) is estimated as a function of this threshold in order to determine the "optimal" components.

3.3 Information Bottleneck Method

The information bottleneck method (IB) of Tishby et al (1999), enables the construction of a compact representation for each class, which maintains its most relevant features. In Rate Distortion theory a quantitative measure for the quality of a compact representation is provided by a *distortion function*. In general, definition of this function depends on the application: in speech processing, the relevant acoustic distortion measure is rather unknown, since it is a complex function of perceptual and linguistic variables [53]. IB method provides an information theoretic formulation and solution to the tradeoff between compactness and quality of a signal's representation [129, 122, 53]. In the supervised learning framework, features are regarded as relevant if they provide information about a target. IB method assumes that this additional variable y (the target) is available. In the case of speech processing systems, the available tagging y of the audio signal (as speech / non speech class, speakers or phonemes) guides the selection of features during training. The relevance of information in the representation of an audio signal, denoted by x, is defined as the amount of information it holds about the other variable y. If we have an estimate of their joint distribution $p(x, y)$, a natural measure for the amount of relevant

information in x about y is given by Shannon's mutual information between these two variables:

$$I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.8}$$

where the discrete random variables $x \in X$ and $y \in Y$ are distributed according to $p(x)$, and $p(y)$, respectively. Further, let $\tilde{x} \in \tilde{X}$ be another random variable which denotes the compressed representation of x; x is transformed to $\tilde{x}$ by a (stochastic) mapping $p(\tilde{x}|x)$. Our aim is to find an $\tilde{x}$ that compresses x through minimization of $I(\tilde{x};x)$, i.e. the mutual information between the compressed and the original variable. At the same time, the compression of the resulting representation $\tilde{x}$ should be minimal *under the constraint* that the relevant information in $\tilde{x}$ about y, $I(\tilde{x};y)$ stays above a certain level. This constrained optimization problem can be expressed via Lagrange multipliers, with the minimization of the *IB variational functional*:

$$\mathscr{L}\left\{p(\tilde{x}|x)\right\} = I(\tilde{x};x) - \beta I(\tilde{x};y) \tag{3.9}$$

where β, the positive Lagrange multiplier, controls the tradeoff between compression and relevance. The solution to this constrained optimization problem has yielded various iterative algorithms that converge to a reduced representation $\tilde{x}$, given $p(x,y)$ and β [122]. The *sequential optimization algorithm* (sIB) yields a fixed number of hard clusters as output. The input consists of the joint distribution $p(x,y)$, the tradeoff parameter β and the number of clusters $M = |\tilde{X}|$. During initialization, the algorithm creates a random partition $\tilde{X}$, i.e. each element $x \in X$ is randomly assigned to one of the M clusters $\tilde{x}$. Afterwards, the algorithm enters an iteration loop. At each iteration step, it cycles through all $x \in X$ and tries to assign them to a different cluster $\tilde{x}$ in order to *increase* the IB functional:

$$\mathscr{L}_{max} = I(\tilde{x};y) - \beta^{-1} I(\tilde{x};x) \tag{3.10}$$

This is equivalent to minimization of the functional defined in 3.9, and it is used for consistency with [122]. The algorithm terminates when the partition does not change during one iteration. This is guaranteed because $\mathscr{L}_{max}$ is always upper bounded by some finite value. To prevent the convergence of the algorithm to a local maximum (i.e., a suboptimal solution), several runs with different initial random partitions need to be performed [122].

3.3.1 Application to Multi-Scale Spectro-temporal Modulations

We estimated the power distribution in the spectro-temporal modulations of speech signals, and compared it to the modulation statistics of other sounds. The auditory model of Shamma et al [137] was the basis for these estimations. The feature tensor $\mathcal{Z}$ represents a discrete set of *continuous* features in the frequency (F), rate (R) and scale (S) subspaces: $z_{i_1,i_2,i_3} = \mathcal{Z}_{i_1,i_2,i_3} \in \mathbb{R}^{+F\times R\times S}$. Since each response z_{i_1,i_2,i_3} is collected over a time frame, it can be interpreted as the average count of an inherent binary event (in the case of a neural classifier, this would be a spike). We therefore consider each response at *location* indexed by i_1, i_2, and i_3, as a binary feature whose number of occurences in a time interval is represented by z_{i_1,i_2,i_3}.
Let the location of a response be denoted by $c_i,\ i = 1,\ldots,F\times R\times S$, such that $z_{i_1,i_2,i_3} = z_{c_i}$. The 3 - dimensional modulation spectrum (frequency - rate - scale) is divided then into $F \times R \times S$ bins centered at $(f_{i_1}, r_{i_2}, s_{i_3})$. Given a training list of M feature tensors $Z^{(k)}$, $k = 1,\ldots,M$ and its corresponding targets $y^{(j)}$, $j = 1, 2$ (speech - nonspeech tags), we can now build a count matrix $K(c, y)$ which indicates the frequency of occupancy of the i^{th} discrete subdivision of the modulation spectrum in the presence of a certain target value y. Normalizing this count matrix such that its elements sum to 1, provides an estimate of the joint distribution $p(c, y)$, which is all the IB framework requires. We assume that M is large enough such that the estimate of $p(c, y)$ is reliable, although it has been reported that satisfactory results were achieved even in cases of extreme undersampling [122].
For the purpose of speech - nonspeech discrimination, the target variable y has only two possible values, y_1 and y_2. We choose to cluster the features c into 3 groups, one composed of features relevant to y_1, the second of features relevant to y_2, whereas the third cluster includes features that are not relevant for a specific target. Let us denote a compressed representation (a reduced feature set) by t and the deterministic mapping obtained by sIB algorithm as $p(t|c)$. We discard the cluster t_j whose contribution :

$$C_{I(t;y)}(t_j) = \sum_y p(t_j, y) \log \frac{p(t_j, y)}{p(t_j)p(y)} \tag{3.11}$$

to $I(t, y)$ is minimal, because its features are mostly irrelevant in this case. Therefore, we don't even have to estimate the responses at these locations of the modulation spectrum. This implies an important reduction in computational load, still keeping the maximally informative features with respect to the task of speech-nonspeech discrimination. To find out the identity of the

remaining two clusters, we compute:

$$p(t,y) = \sum_c p(c,y)p(t|c) \tag{3.12}$$

$$p(t) = \sum_y p(t,y) \tag{3.13}$$

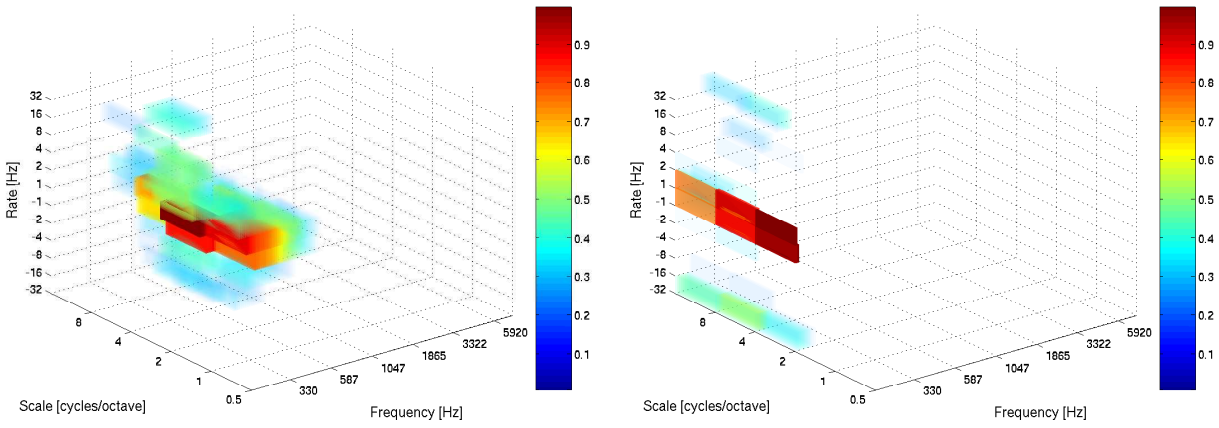
$$p(y|t) = \frac{p(t,y)}{p(t)} \tag{3.14}$$

The cluster that maximizes the likelihood $p(y_1|t)$ contains all relevant features for y_1; the other for y_2. We denote, hence, the first cluster as t_1 and the latter as t_2. The typical pattern (3-dimensional distribution) of features relevant for y_1 is given by $p(c|t = t_1)$, while for y_2 is given by $p(c|t = t_2)$. According to Bayes rule, these are defined as:

$$p(c|t = t_j) = \frac{p(t = t_j|c)p(c)}{p(t = t_j)}, \qquad j = 1, 2 \tag{3.15}$$

Figure 3.1b presents an example of the relevant modulation spectrum of each sound ensemble, speech and non-speech (music, animal sounds and various noises). Speech examples were taken from the TIMIT Acoustic-Phonetic Continuous Speech Corpus. Music examples were selected from the authors' music collection. Animal vocalizations consist of bird sounds and were taken from [124]. The noise examples (taken from Noisex) consist of background speech babble in locations such as restaurants and railway stations, machinery noise and noisy recordings inside cars and planes. Training set consists of 500 speech and 560 non-speech samples. One single frame of 500ms is extracted from each example, starting at a certain sample offset in order to skip initial periods of silence.

In some sense, Figure 3.1 presents the statistical structure of the modulation spectrum of each sound ensemble. Speech class (Figure 3.1a) is more homogeneous since it consists exclusively of TIMIT samples. It is characterized by a triangular-like structure corresponding to the pitch of the voices and their harmonics; due to the logarithmic frequency axis (in octaves), an upward change in scale is matched to the same increase in frequency band. The harmonic structure due to voiced speech segments is mainly depicted at the higher spectral modulations (2-6 cycles/octave). Scales lower than 2 cycles/octave represent the spectral envelope or formants [21]. Temporal modulations in speech spectrograms are the spectral components of the time trajectory of spectral envelope of speech. They are dominated by the syllabic rate of speech, typically close to 4 Hz, whereas most relevant temporal modulations are below 8 Hz in the figure. It can also be noticed



(a) (b)

Figure 3.1: $p(c|t = t_1)$ for speech (a) and $p(c|t = t_2)$ for non-speech class (b). Cluster t_1 holds 37.5% and t_2 holds 24.7% of all responses. The remaining 37.8% are irrelevant.

that the lower frequencies - between 330 and 1047 Hz - are more prominent than higher ones, in accordance to the analysis in [140], due to the dominance of voice pitch over these lower frequency bands [21].

Non-speech class (Figure 3.1b) consists of quite dissimilar sounds - natural and artificial ones. Therefore, its modulation spectrum has quite "flat" structure, representing bins not present in speech modulation spectrum: rates lower than 2 Hz in combination with frequencies lower than 330 Hz and scales less than 1 cycle/octave; frequency-scale distribution hasn't any structure as in the case of speech.

Knowledge of such compact modulation patterns allows us to classify new incoming sounds based on the similarity of their cortical-like representation (the feature tensor $\mathcal{Z}$) to the typical pattern $p(c|t = t_1)$ or $p(c|t = t_2)$ [138]. In [91] we assessed the similarity (or correlation) of $\mathcal{Z}$ to $p(c|t = t_1)$ or $p(c|t = t_2)$, by the spectro-temporal modulation index (STMI) which is used to the assessment of speech intelligibility [34]. We adapted STMI to handle the contribution of different frequency bands as follows: the spectro-temporal modulation index (STMI) [34], defined below between corresponding Ω, ω *and* f channels:

$$\rho_s(\Omega, \omega, f) = \sqrt{\frac{1}{1 + \left(\frac{\mathcal{Z}(\Omega,\omega,f) - p(x|\tilde{x}=\tilde{x}_2)}{\sigma_{\mathcal{Z}}(\Omega,\omega,f)}\right)^2}} \tag{3.16}$$

where $\sigma_{\mathcal{Z}}(\Omega, \omega, f)$ is the standard deviation of the auditory representation extracted over a fixed

time-frame (500 ms) at each channel, whereas $\mathcal{Z}(\Omega, \omega, f)$ is the corresponding mean value.

Finally we derived the average of $\rho_s(\Omega, \omega, f)$ over the channels $(\Omega, \omega, f) \in \tilde{X}_2$, i.e., over which $p(x|\tilde{x} = \tilde{x}_2) \neq 0$:

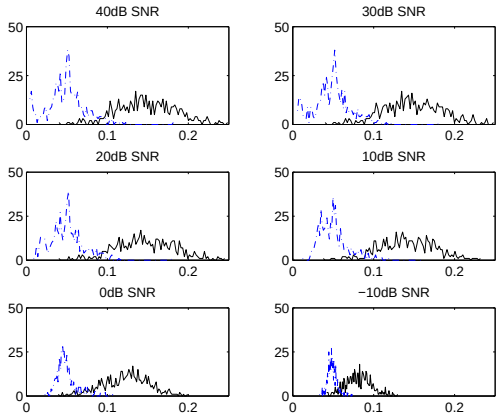
$$\rho(\mathcal{Z}) = \frac{1}{|\tilde{X}_2|} \sum_{(\Omega,\omega,f)\in\tilde{X}_2} \rho_s(\Omega, \omega, f). \tag{3.17}$$

Moreover, we took into account the similarity of audio signals to both cluster "prototypes" (speech and non-speech),$p(x|\tilde{x} = \tilde{x}_1)$ and $p(x|\tilde{x} = \tilde{x}_2)$, ρ_1 and ρ_2 respectively, by taking their ratio:

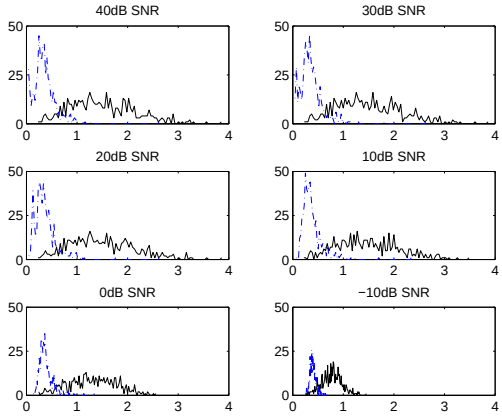
$$R(\mathcal{Z}) = \frac{\rho_2}{\rho_1} \tag{3.18}$$

We calculated the STMI (ρ) and corresponding ratio (R) for all training examples and noise conditions. Figure 3.2 shows the histograms of ρ and R computed on speech (continuous curve) and non-speech examples (dashed curve). The histograms form two distinct clusters with a small degree of overlap in the case of ρ, whereas decision threshold depends on the SNR condition especially for low SNR (0dB, -10dB). In the case of R distribution the overlap is increased, however the decision threshold is less sensitive to the variation of SNR. This trend is reflected in the results in the benchmark test presented below. A simple threshold check was then used for discriminating speech from non-speech events. Threshold setting implies a trade-off between the false acceptance rate (FAR) and false rejection rate (FRR) for each class. In this case, we have set threshold at a fixed value θ that minimizes the total number of segments incorrectly assigned to each class at the highest SNR level (40dB).

The system was compared to the system in [95] which was based on the same auditory features but used a multilinear dimensionality reduction technique - Higher Order Singular Value Decomposition (HOSVD) [68] - and Support Vector Machines (SVMs) for classification. We evaluated systems performance in voice activity detection under varying noise conditions, using F-measure. In low levels of additive noise, our system was almost equivalent to the system of [95], whereas in low SNR conditions the proposed method exhibited superior performance (SNR$\leq$ 0dB). For evaluation purposes, we have also implemented another segment-based system based on MFCCs and Zero Crossing Rates (ZCRs); these features were also extracted on a frame basis and their mean values in each segment were given as input to an SVM classifier. This system was used as a reference system to show the robustness of the auditory features to various noise conditions.



(a)



(b)

Figure 3.2: *Histogram of STMIs ρ (a) and ratios of STMIs R (b) computed on nonspeech (dashed) and speech examples (continuous curve).*

3.4 Relevance - Redundancy trade-off

3.4.1 Feature selection based on Maximum Relevance and Min-Redundancy

The *maximal relevance* (MaxRel) feature selection criterion simply selects the features most relevant to the target class c [101]. Relevance is usually defined as the mutual information $I(x_j; c)$ between feature x_j and class c. Through a sequential search which does not require estimation of multivariate densities, the top m features in the descent ordering of $I(x_j; c)$ are selected [101]; the cross-validation classification error for an increasing number of these sequential features needs to be computed, in order to determine the optimal size of feature set, m.

Still, the "m best features" are not necessarily the "best m features" due to the redundancy

among them. The purpose of feature selection is to find a feature set S_m with m features $\{x_i\}$ which *jointly* have the largest dependency on the target class c. This is the Max-Dependency scheme [101]:

$$\max D(S,c),\ D = I(\{x_i, i = 1,\ldots,m\}; c) \tag{3.19}$$

with the mutual information $I(S_m; c)$ defined as:

$$I(S_m; c) = \int \ldots \int p(x_1,\ldots,x_m,c) \log \frac{p(x_1,\ldots,x_m,c)}{p(x_1,\ldots,x_m)p(c)} \tag{3.20}$$

It has been shown through theoretical analysis in [101] that Max-Dependency is equivalent to the "minimal-Redundancy-Maximal-Relevance" (mRMR) criterion, which spots near-optimal features for classification through an incremental search method optimizing the following condition:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right] \tag{3.21}$$

where $I(x_j; x_i)$ is the mutual information between features x_j and x_i, i.e., redundancy, and S_{m-1} is the initially given set of $m-1$ features. The m^{th} feature selected from the set $X - S_{m-1}$ maximizes relevance *and* reduces redundancy. The computational complexity of both incremental search methods, MaxRel and mRMR, is $O(|S|M)$ [101]. However it is necessary to use heuristics during training for discovering the optimal relation between relevance and redundancy of features.

3.4.2 Most Relevant of Least Redundant Features

HOSVD efficiently addresses the differing degrees of redundancy between the features in each subspace by selecting mutually independent features. When the variables are independent of each other, their joint probability density function can be approximated by the product of the marginal probability densities:

$$p(x_1,\ldots,x_m) = p(x_1)\ldots p(x_m) \tag{3.22}$$

and the mutual information between features, or redundancy $J(S_m) = J(x_1,\ldots,x_m)$, is minimized:

$$J(x_1,\ldots,x_m) = \int \ldots \int p(x_1,\ldots,x_m) \log \frac{p(x_1,\ldots,x_m)}{p(x_1)\ldots p(x_m)} dx_1 \ldots dx_m \approx 0. \tag{3.23}$$

This is clearly demonstrated in Fig. 3.3, where the distribution of Mutual Information (MI) $I(x_j; x_i)$ *between* features is depicted before (red triangles) and after applying HOSVD (yellow triangles); MI between pairs of "packed" features is significantly smaller (almost zero) than MI between original features. This figure was produced using the dataset from MEEI (cf Chapter 5, on voice pathology detection). Another advantage of using HOSVD first, is the noise reduction

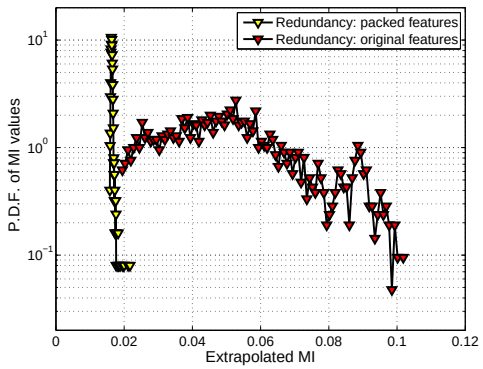


Figure 3.3: Redundancy reduction through HOSVD is depicted as the distribution of Mutual Information $I(x_j; x_i)$ *between* pairs of features before (red triangles) and after applying HOSVD (yellow triangles).

inherent to the SVD process.

Since the redundancy $J(S_m) \approx 0$ (due to equation 3.22), the mutual information $I(S_m; c)$ in the Max-Dependency scheme (equation 3.19) reduces to $J(S_m, c)$:

$$J(S_m, c) = \int \dots \int p(x_1, \dots, x_m, c) \log \frac{p(x_1, \dots, x_m, c)}{p(x_1) \dots p(x_m) p(c)} dx_1 \dots dx_m dc. \tag{3.24}$$

The term $J(S_m, c)$ attains its maximum value when all variables are maximally dependent; if one feature is selected at one time, the others being already selected ("first-order" incremental search [101]), then this feature should be maximally dependent on the class variable. This is the Max-Relevance criterion [101]; for the "first-order" incremental search then, the Max-Dependency scheme reduces to Max-Relevance when redundancy of features has been minimized before. In effect, the joint effects of features on the target class need not to be considered; and through Max-Relevance we avoid the estimation of multivariate densities which is more difficult and less accurate than calculating bivariate densities.

We tested both maximum relevance and max relevance-min redundancy approaches in order to select n sequential feature sets $S_1 \subset \dots \subset S_k \subset \dots \subset S_n$ and computed the respective error using SVM classifier and the same data set. Figure 3.4 presents the results in terms of

equal error rate (EER) in detecting speech from nonspeech; these results were also presented in [82]. We deduce that the max-relevance (MaxRel) is slightly better than max-relevance plus min-redundancy (mRMR) approach. The reason that mRMR doesn't offer any advantage over MaxRel - it is actually worse - is that features redundancy has been already reduced by using HOSVD as a first step. Redundancy is not exactly zero because the truncation of the least significant singular values destroys the orthogonality of the singular vectors (in contrast to the SVD case) [68]. In order to determine the least redundant features then, we could compare their relevance to the classification task, with the average mutual information between pairs of features (mean pairwise redundancy). The lowest singular values in every mode for which mean redundancy is comparable to their relevance, could be safely discarded.

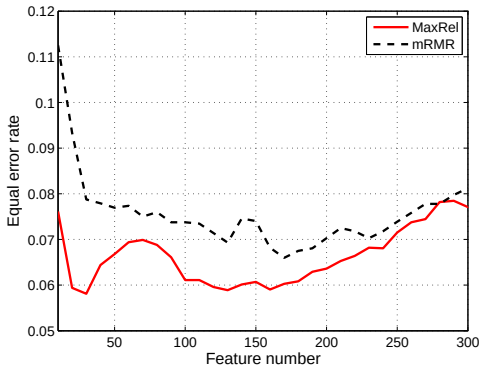


Figure 3.4: SVM classifier equal error rate using mRMR and MaxRel features for speech/nonspeech discrimination on broadcast news.

Both redundancy and relevance of features need to be addressed before selecting points in the feature space for an efficient, robust automatic classification system. Depending on the task, data available for training (noisy or not) and features extracted, a different relevance/redundancy trade-off occurs. In the next chapters, this trade-off is experimentally assessed in different audio classification tasks.

Chapter 4

Speech Discrimination based on Modulation Spectra

Synopsis of the Chapter

In audio content analysis, the discrimination of speech and non-speech is the first processing step before speaker segmentation and recognition, or speech transcription. Speech/non-speech segmentation algorithms usually consist of a frame based scoring phase using MFCC features, combined with a smoothing phase. In this Chapter, a content based speech discrimination algorithm is designed to exploit long-term information inherent in modulation spectrum (presented in Chapter 2). In order to address the varying degrees of redundancy and discriminative power of the acoustic and modulation frequency subspaces, we first employ a generalization of SVD to tensors (Higher Order SVD) to reduce dimensions. Projection of modulation spectral features on the principal axes with the higher energy in each subspace results in a compact set of features with minimum redundancy. We further estimate the relevance of these projections to speech discrimination based on mutual information to the target class. This system is built upon a segment based SVM classifier in order to recognize the presence of voice activity in audio signal. Detection experiments using Greek and U.S. English broadcast news data composed of many speakers in various acoustic conditions suggest that the system provides complementary information to state-of-the-art mel-cepstral features.

This chapter is based upon the following publications:

- Markaki M. and Stylianou Y., *Discrimination of Speech from nonspeech in broadcast news based on modulation frequency features*, ISCA Tutorial and Research Workshop: Speech Analysis and Processing for Knowledge Discovery, 2008.
- Markaki M. and Stylianou Y., *Dimensionality Reduction of Modulation Frequency Features for Speech Discrimination*, InterSpeech, 2008.
- Markaki M., Holzapfel A. and Stylianou Y., *Singing Voice Detection using Modulation Frequency Features*, ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, 2008.
- Markaki M. and Stylianou Y., *Evaluation of Modulation Frequency Features for Speaker Verification and Identification*, EUSIPCO, 2009.
- Markaki M. and Stylianou Y., *Discrimination of Speech from Nonspeeech in Broadcast News Based on Modulation Frequency Features*, Speech Communication doi:10.1016/j.specom.2010.08.007, 2010.

Many thanks to Andre Holzapfel for the labeled data (with greek rebetico music) for the singing voice detection.

4.1 Introduction

The increasingly larger volumes of audio that are amassing nowadays, require a pre-processing in order to remove information-less content before storing. Usually the first stage of processing partitions the signal into primary components such as speech, and non-speech before speaker segmentation and recognition, or speech transcription.

Reviewing relevant past work, many approaches in the literature have examined various features and classifiers. In telephone speech adaptive methods such as short-term energy-based methods, first measure the energy of each frame in the file and then set the speech detection threshold relative to the maximum energy level. A simple energy level detector that is very efficient in high signal-to-noise ratio (SNR) conditions would fail in lower SNR or when music and noise are present (which also contain substantial energy). In [118] a real-time speech/music classification system was presented based on zero-crossing rate and short-term energy over a 2.4 sec segment of broadcast FM radio. Scheirer and Slaney [119] proposed another real-time speech/music discriminator using thirteen features in time, frequency and cepstrum domain for modeling speech and music and different classification schemes over 2.4 sec segments. Methods based on such low level perceptual features are considered less efficient when a window smaller than 2.4 sec is used, or when more audio classes such as environmental sounds are taken into account [70].

Mel-frequency cepstral coefficients (MFCC) - the most commonly used features in speech and speaker recognition systems - have been successfully applied in audio indexing task [5, 14, 70]. For applications in which the audio is also transcribed, these features are available at no additional computational cost for direct audio search. Each audio frame can be represented with either just the "static" cepstra or also augmenting the representation with the first and second order time derivatives to capture dynamic features in the audio stream. It has been extensively documented that it is difficult to accurately discriminate speech from nonspeech given a single frame [5, 70, 96]. Speech/non-speech segmentation algorithms usually consist of a frame based scoring phase using MFCC features, combined with a smoothing phase. The general approach used for audio segmentation is based on Maximum Likelihood (ML) classification of a frame with Gaussian mixture models (GMMs) using MFCC features [14]. The smoothing of likelihoods, when using the GMM framework, assumes that the feature vectors of neighboring frames are independent given a certain class; this smoothing is commonly applied by the GMM-based algorithms either for speech-nonspeech and audio classification or for speaker recognition [14, 113]. In [63], SVM

classifier was used based on cepstral features; median smoothing of SVM output scores over 1 sec segments improved frame-based classification accuracy by $\sim 30\%$. The performance of SVM-based system on different domains was more consistent or even better than GMMs based on the same cepstral features [63].

In [70, 125, 5], the classification entity is a sequence of frames (a segment) rather than a single frame. In [70, 125], segments were parameterized by the mean value and standard deviation of frame-based features over a much longer window.Audio classification was performed using SVMs in [70], and GMMs in [125]. In [5], a segment based classifier was built unifying both frame based scoring phase and the smoothing phase. Audio segments were modeled as supervectors through a segment based generative model and each class (speech, silence, music) was modeled by a distribution over the supervector space. Classification of speech/non-speech classes proceeded then using either GMMs or SVMs [5].

In this work we first compare and then combine the speech discrimination ability of cepstral features to that of modulation spectral features [50, 8]. Dynamic information provided by the modulation spectrum captures fast and slower time-varying quantities such as pitch, phonetic and syllabic rates of speech, tempo of music, etc [50, 8]. In [108], it was suggested that these high level modulation features could be combined with standard mel-cepstral features to enhance speaker recognition performance. Hence these features could be available at no additional computational cost for direct audio search (as MFCC).

Still, the use of modulation spectral features for pattern classification is prevented by their dimensionality. Methods addressing this problem have proposed critical band filtering to reduce acoustic frequencies, and a continuous wavelet transform instead of a Fourier transform [126], or a discrete cosine transform [64] for modulation frequencies. In [108], dimensionality reduction was performed either by averaging across modulation filters or across acoustic frequency bands.

We adopt a different approach towards dimensionality reduction of this two-dimensional representation. We employ a higher order generalization of singular value decomposition (HOSVD) to tensors [68], and retain the singular vectors of acoustic and modulation frequency subspaces with the higher energy. Joint acoustic and modulation frequencies are projected on the retained singular vectors in each subspace to obtain the multilinear principal components (PCs) of the sound samples. In this way the varying degrees of redundancy of the acoustic and modulation frequency subspaces are efficiently addressed. This technique has been successfully applied in auditory-based features with multiple scales of time and spectral resolution in [96].

Truncation of singular vectors based on their energy addresses features redundancy; to assess

their discriminative power, we need an estimate of their mutual information (MI) to the target class (speech versus non-speech, i.e., noise, music, speech babble) [22]. By first projecting the high-dimensional data to a lower order manifold, we can approximate the statistical dependence of these projections to the class variable with reduced computational effort. We spot near-optimal PCs for classification among those contributing more than an energy threshold through an incremental search method based on mutual information [101].

In Chapter 2, we overviewed the modulation frequency analysis framework which is commonly used [8]. The multilinear dimensionality reduction method and the mutual information-based feature selection were presented in Chapter 3. In the same Chapter we also discussed the practical implementation of mutual information estimation based on the joint probability density function for two variables and its marginals.

In Section 4.1 then, we describe the experimental setup, the databases, the classification systems and the results using the proposed features, mel cepstral features and the concatenation of both feature sets. Finally, in Section 4.5 we present our conclusions.

4.2 Data Collection

We first tested the methods described in Chapter 3 on audio data recorded from broadcasts of Greek TV programs (ERT3). The database was manually segmented and labeled at CSD. The labeled dataset used in these experiments consists of 6 hours; it is available upon request from the first author.

Audio data are all mono channel and 16 bit per sample, with 16 kHz sampling frequency. Speech data consists of broadcast news and TV shows recorded in different conditions such as studios or outdoors, under quiet conditions or with background noise; also, some of the speech data have been transmitted over telephone channels. Non-speech data consists of music (mainly audio signals at the beginning and the end of TV shows, or music accompanying talks of political candidates), outdoors noise from moving cars, beeps, crowd, claps, or very noisy unintelligible speech due to many speakers talking simultaneously (speech babble). We used 7 broadcast shows for training, with minimum duration of $\sim$ 6 min, and maximum duration of $\sim$ 1 hour (1 and a half hour in total). Fifteen shows were used for testing with minimum duration of $\sim$ 6 min and maximum duration of $\sim$ 1 hour ($\sim$ 4 and a half hours in total). Each file was partitioned into 500 ms segments for long-term feature analysis. We extracted evenly spaced overlapping segments every 250 ms for speech and every 50 ms for non-speech (in order to maximize non-speech data).

We also conducted experiments on the NIST RT-03 evaluation data distributed by LDC (LDC2007S10). The dataset we used consisted of six audio files with 30 minutes duration each, recorded in February 2001 from U.S. radio or TV broadcast news shows, from ABC, CNN, NBC, PRI, and VOA. For parameter tuning, we performed 5-fold cross-validation experiments on a subset of ~ 1 hour of this data; system performance was evaluated on the rest of data.

4.3 Methods

4.3.1 Feature Extraction and Classification

The modulation spectrogram was calculated using Modulation Toolbox [7]. For every 500 ms block modulation spectrum features were generated using a 128 point spectrogram with a Gaussian window. The envelope in each subband was detected by a magnitude square operator. To reduce the interference of large dc components of the subband envelope, the mean was subtracted before modulation frequency estimation. One uniform modulation frequency vector was produced in each one of the 65 subbands. Due to a window shift of 32 samples, each modulation frequency vector consisted of 125 elements up to 250 Hz. Feature calculation runtime is $O(N \log_2 N)$, since the estimation of modulation spectral features consists of two FFTs.

The mean value was computed over the training set and subtracted from all matrices; stacking of the training matrices produced the data tensor $\mathcal{D} \in \mathbb{R}^{65\times125\times7200}$. The singular matrices $\mathbf{U}^{(1)} \equiv \mathbf{U}_{freq} \in \mathbb{R}^{65\times65}$ and $\mathbf{U}^{(2)} \equiv \mathbf{U}_{mod} \in \mathbb{R}^{125\times125}$ were directly obtained by SVD of the "matrix unfoldings" $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$ of $\mathcal{D}$ respectively. By retaining the singular vectors that exceeded a contribution threshold of 1% in each mode (eq. 3.4), resulted in the truncated singular matrices $\hat{U}_{freq} \in \mathbb{R}^{65\times24}$ and $\hat{U}_{mod} \in \mathbb{R}^{125\times29}$. Features were projected on $\hat{U}_{freq}$ and $\hat{U}_{mod}$ according to eq. (3.6) resulting in matrices $\mathbf{Z} \in \mathbb{R}^{24\times29}$; these were subsequently reshaped into vectors before MI estimation, feature selection and SVM classification. All features were normalized by their corresponding standard deviation estimated from the entire training set to reduce their dynamic range before classification (their mean value has already been set to zero before projecting them to the truncated singular matrices).

HOSVD is the most costly process in our system but it is performed only once. HOSVD consists of the SVD of two data matrices $N \times k$ each composed of N k-dimensional vectors; computational complexity of SVD transform is $O(Nk^2)$. N is either the acoustic frequency dimension or the modulation frequency dimension; respectively, k is the product of the modulation or the acoustic frequency dimension multiplied by the size of the training dataset.

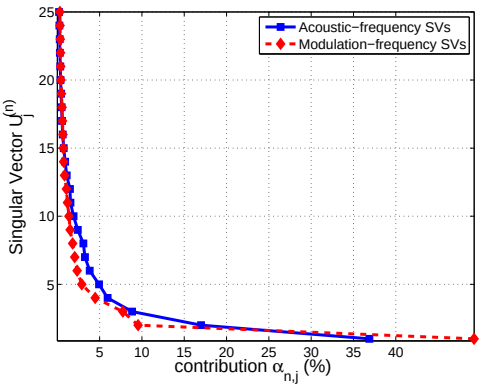


Figure 4.1: Contribution $\alpha_{n,j}$ of the first 25 singular vectors (SVs) $U_j^{(1)}$, $U_j^{(2)}$, $j = 1, \ldots, 25$, to the acoustic and modulation frequency subspaces, respectively.

Figure (4.1) presents the contribution of the first 25 singular vectors $U_j^{(1)}$ and $U_j^{(2)}$, $j = 1, \ldots, 25$, in the acoustic and modulation frequency subspaces, respectively. Ordering of the $n-$mode singular values $\lambda_{n,j}$ implies that the "energy" of modulation spectral representation is concentrated in the lower j-indices. In addition, Figure (4.1) shows that variance in the acoustic frequency subspace slightly exceeds that in the modulation frequency subspace; rather more acoustic frequency SVs should be retained for "best rank approximation" of a modulation spectral representation.

For the data discretization involved in MI estimation, the number of discrete bins along each axis was set to $b^* = 8$ according to the procedure described in [123]. Figure 4.2 compares the relevance of features in the original and reduced representation. The number of relevant features in the original representation is large, posing a serious drawback to any classifier: 1147 out of the 8125 features (14.12%) have mutual information to the target class more than 0.04 bits. As Figure 4.2a depicts, the most relevant among the original features are mainly distributed along the modulation frequency axis: they span the ranges of the lower syllabic and phonetic rates of speech ($\sim 4 - 30$ Hz) as well as the range of pitch of the majority of speakers, i.e., up to ~ 200 Hz). They also appear confined to the lower acoustic frequency bands up to ~ 2500 Hz.

The HOSVD redundancy reduction method has reduced dimensions in each subspace separately. Therefore, the differential relevance of the two subspaces is preserved in the compressed representation as MI estimation reveals. Figure (4.2b) presents MI estimates between each of the first 25 singular vectors and the speech/non-speech class variable for the training set. The subspace spanned by the first two acoustic frequency singular vectors (SVs) and the first 15 modulation frequency SVs appear to be the most relevant to speech-non-speech discrimination with

(a) (b)

Figure 4.2: Relevance of the original and compressed modulation spectral features: (a) Mutual information (MI) between the acoustic and modulation frequencies (65×125 dimensions) and the speech/non-speech class variable. (b) MI between the first 25 singular vectors in each subspace and the speech/non-speech class variable.

much lower peaks elsewhere. According to MI criterion, then, variance in modulation frequency subspace is more relevant to the classification task. In addition, the number of relevant features is significantly reduced in the compressed representation: only 27 out of the 696 "packed" features (3.94%) have mutual information to the target class more than 0.04 bits. Still the maximum value of relevance to the classification task is increased.

4.4 Results

Classification of segments was performed using support vector machines. SVMs find the optimal boundary that separates two classes maximizing the margin between separating boundary and closest samples to it (support vectors) [60]. We have used SVMlight [60] with a Radial-Basis-Functions kernel.

We evaluate system performance on the validation and the test set using the Detection Error Trade-off curve (DET) [94]. The DET curves depict the false rejection rate (or miss probability) of the speech detector versus its false acceptance rate (or false alarm probability). DET curves are quite similar to the Receiver Operating Characteristic (ROC) curves, except that the detection error probabilities are plotted on a nonlinear scale. This scale transforms the error probabilities by mapping them to the corresponding Gaussian deviates. Thus DET curves are straight lines when the underlying distributions are Gaussian. This makes DET plots more intelligible than ROC plots [94]. We have used the matlab files that NIST has made available for producing DET curves with the matlab software package [94].

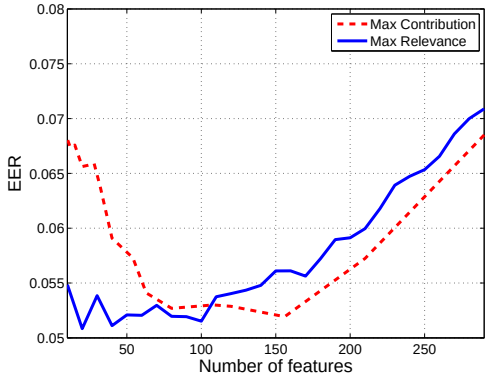


Figure 4.3: SVM classifier equal error rate (EER) as a function of features selected in terms of relevance or contribution.

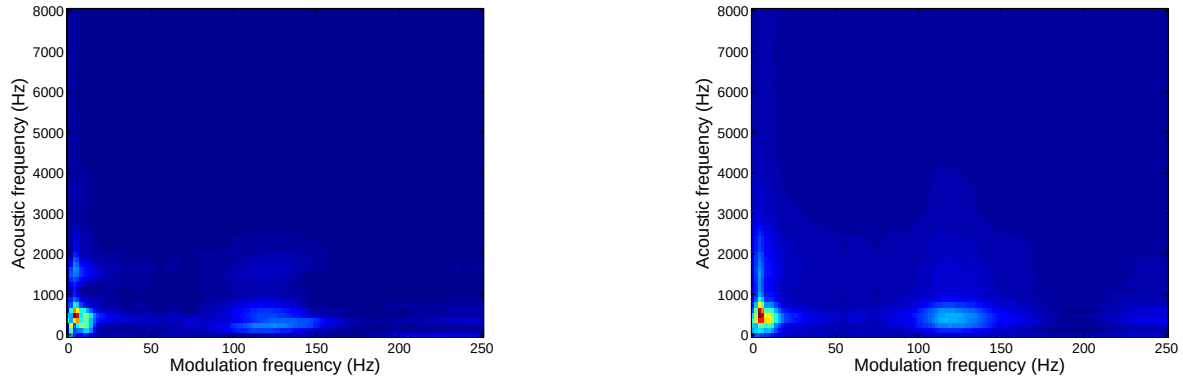
Since the costs of miss and false alarm probabilities are considered equally important, the minimum value of the detection cost function, DCF_{opt}, is:

$$DCF_{opt} = \min\left(P_{miss} * P_{speech} + P_{false} * P_{non-speech}\right). \tag{4.1}$$

where P_{speech} and $P_{non-speech}$ are the prior probabilities of speech and non-speech class respectively. We also report the equal-error rate (EER) - the point of DET curve where the false alarm probability equals the miss probability.

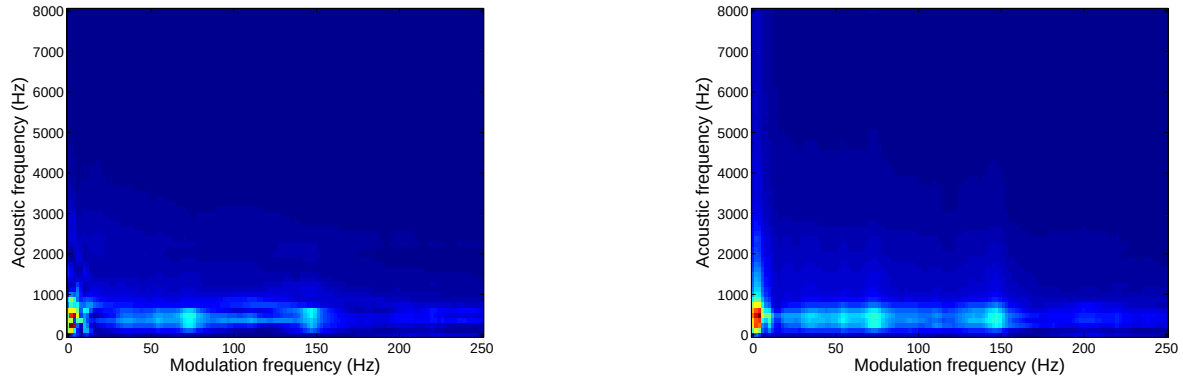
In Figure 4.3 we compare the SVM classifier EER on the validation data set when using features selected either in terms of contribution or relevance. According to the maximum contribution criterion, we retained singular vectors with contributions varying between 0.5% up to 6% (eq. 3.4). The dimensionality of the reduced features varied between $18 \times 18 = 324$ dimensions up to $3 \times 3 = 9$ dimensions, respectively. EER was lowest for the configuration of $13 \times 12 = 156$ dimensions; increase in dimensionality beyond 156 features induced poor generalization whereas for less than $9 \times 6 = 54$ features, the performance became progressively worse. Under the maximum relevance selection criterion, just 21 features yielded the best classification performance in terms of EER.

Figures 4.4, 4.5, 4.6 depict the rank$-(13, 12)$ approximation of modulation spectra (eq. 3.6) as well as their reconstruction from the 21 most relevant features for speech, music and noise signals, respectively. Energy at modulations that characterize speech at the lower acoustic frequency bands, corresponding to syllable and phonemic rates (< 40 Hz) and the pitch of speaker, remain prominent in both representations of speech (Fig. 4.4). In Fig. 4.5, the energy at modulations



(a) (b)

Figure 4.4: (a) Rank$-(13, 12)$ approximation (eq. 3.6) of $|X_l(k, i)|$ for 500 ms of a speech signal. (b) 21 features approximation for the same speech signal. Energy at modulations corresponding to pitch ($\sim$ 120 Hz) and syllabic and phonetic rates ($<$ 40 Hz) remain prominent.



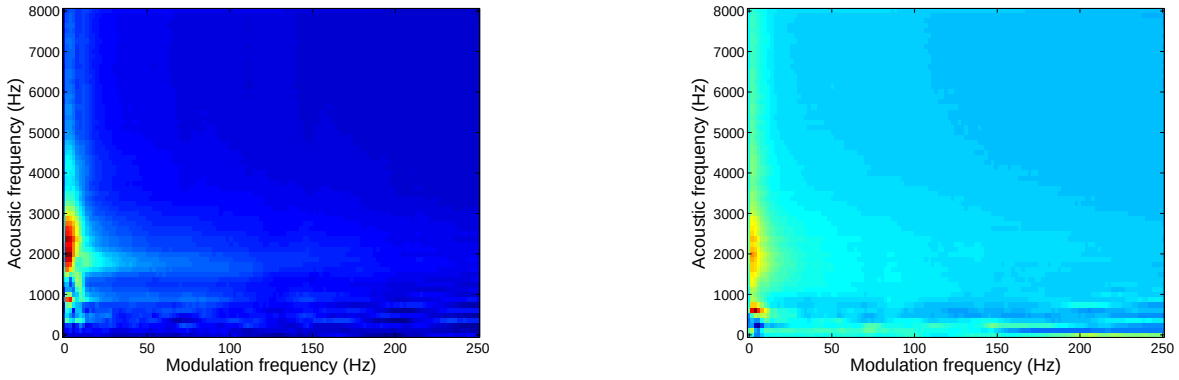
(a) (b)

Figure 4.5: (a) Rank$-(13, 12)$ approximation of $|X_l(k, i)|$ for 500 ms of a music signal. (b) 21 features approximation for the same music signal; the characteristic patterns are not lost.

corresponding to harmonics characterize the music signal (at the beginning of a TV show). The approximate representations of the noise signal (claps and crowd noise outdoors) in Fig. 4.6, depict most of its energy localized in higher frequency bands, and concentrated in lower modulation frequencies.

4.4.1 Combining Modulation and Cepstral Features

Speech/Non-Speech discrimination systems for broadcast news are typically based on the mel-frequency cepstral coefficients that are also routinely used in speech and speaker recognition



(a) (b)

Figure 4.6: (a) Rank$-(13, 12)$ approximation of $|X_l(k, i)|$ for 500 ms of a noise signal (claps and crowd noise outdoors). (b) 21 features approximation for the same signal.

systems. The features used in the baseline system consist of 12th-order Mel frequency cepstral coefficients (MFCCs), log-energy, along with their first and second differences to capture dynamic features in the audio stream [14]. This makes a frame-based feature vector of 39 elements (13×3) The features were extracted from 30 ms audio frames with a 10 ms frame rate, i.e. every 10 ms the signal was multiplied using a Hamming window of 30 ms duration. Critical-band analysis of the power spectrum with a set of triangular band-pass filters was performed as usual. For each frame, equal-loudness pre-emphasis and cube-root intensity-loudness compression were applied according to Hermansky [54]. The general approach used is maximum-likelihood classification with Gaussian mixture models (GMMs) trained on labeled training data. Still in [63] it was reported that the performance of SVM on different domains was more consistent than GMMs based on the same MFCC features. Therefore, in the subsequent experiments we will use the MFCC-based features with SVM classifiers. This will make easier the comparison between the suggested features and the MFCC-based features. Moreover, we will discuss the fusion of the two sets of features.

In [63], it was found that smoothing the SVM output scores when frame-based features are used, improves the final score in terms of EER (an improvement of about 30% was reported in [63] as compared to the frame-based results prior to smoothing). In [70, 125], segment-based MFCC features were considered. For segments of 500ms, the mean and the standard deviation of 50 frame-based MFCC feature vectors were the segment-based features [70, 125] (i.e., a 78-element feature vector).

We decided to compare the frame-based and segment-based SVM classifiers. We performed

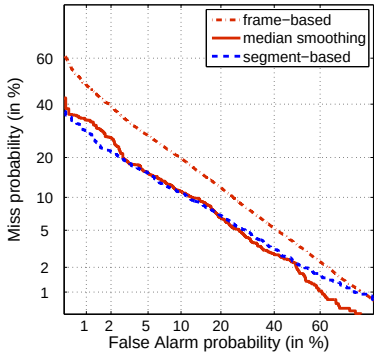


Figure 4.7: Detection Error Trade-off (DET) curves for frame- and segment-based SVM classification using cepstral features, and median smoothing of the frame-level scores; a small subset of training/validation set from the greek broadcast news shows has been used.

2-fold cross-validation on a subset of the Greek training data set (two broadcast shows of total duration 17 minutes, with 26 speakers). Figure 4.7 presents the DET curves for frame-based and segment-based SVM classification results. Applying smoothing, using a median filter, on the frame-based SVM classification results, the frame-based approach is highly improved (solid line in Fig.4.7). Actually it provides on average equivalent result to the segment-based MFCC features. The major disadvantage, however, of any of the frame-based MFCC features approach, is that the computation time for the training and testing of SVM classifier, is much bigger as compared to the segment-based MFCC features. Therefore, we will only consider the segment-based MFCC features for comparison purposes with the suggested modulation spectral features.

Different approaches to information fusion exist [117]: information can be combined prior to the application of any classifier (pre-classification fusion), or after the decisions of the classifier have been obtained (post-classification fusion). Pre-classification fusion refers to feature level fusion in the case of data from a single sensor (such as single channel audio data). When the feature vectors are homogeneous, such as the MFCC features of successive frames of a speech or non-speech audio segment, a single feature vector can be calculated from the mean and standard deviation of the individual feature vectors as in [70, 125]. When different feature extraction algorithms are applied on the input data, the non-homogeneous feature vectors that incur can be concatenated to produce a single feature vector [117]. On the other hand, post-classification fusion can be accomplished either at the matching score level or at the decision level as explained in [59]. According to [59], integration at the feature level is preferable since the features contain richer information about the input data than the matching scores or output decisions of a classifier/matcher. We simply concatenated the different feature vectors into a single representation

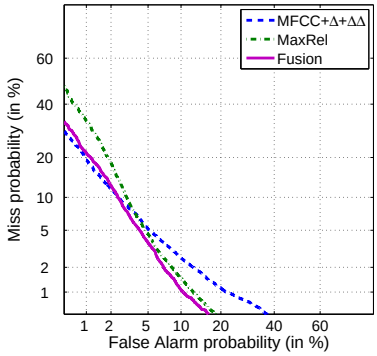


Figure 4.8: DET curves for segment-based SVM classification using cepstral features (MFCC+Δ+ ΔΔ), the 21 most relevant features (MaxRel), and the concatenated feature vector (Fusion) for the same training and testing sets from greek broadcast news shows.

of the input pattern.

Table 4.1: *$\hat{DCF}$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ on test set from Greek shows*

	[13, 12]	MFCCs+Δ + ΔΔ	MaxRel	fusion
EER	5.19	4.79	5.06	4.45
$\hat{DCF}$	5.12	4.63	5.05	4.35
$\hat{P}_{miss}$	4.73	3.20	4.84	2.50
$\hat{P}_{false}$	5.50	6.06	5.27	6.19

Figure 4.8 presents the DET curves and Table 4.1 the respective EER, and the optimal values of $\hat{DCF}$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ for the systems tested using SVM and the same training data set from greek broadcast news shows. MaxRel denotes the system based on the first 21 most relevant features. The last column refers to the fusion of cepstral with MaxRel features; the concatenated (78+21=99)-features vector further reduced $\hat{DCF}$ down to 4.35%. For comparison, we also report the best EER and $\hat{DCF}$ when using the first (R_1, R_2) projections, which were 5.19% and 5.12% respectively for the $[13 \times 12]$ PCs. MaxRel system is better at the low miss probability regions of the DET curve; cepstral features on the other hand yield better classification performance at the low false alarm regions. Fusion of the two feature sets then follows the best of performances across the whole DET curve.

Results on the NIST RT-03 Data

To train our system on US English, we used about 1 hour from U.S. broadcast news from the NIST RT-03 evaluation data (LDC2007S10). Parameter tuning was performed using 5-fold cross-

validation along with SVM classifier. Figure 4.9 presents the SVM classifier equal error rate (EER) as a function of the most relevant modulation spectral features alone, or using them in combination with MFCC features. The EER was minimum when using the 52 most relevant modulation spectral features. On the other hand, using a concatenated feature vector, best performance was achieved through the combination of the **16** most relevant modulation spectral features with MFCC features.Probably, there is some redundancy between modulation spectral features and the augmented MFCC parameters (when Δ and $\Delta\Delta$ are included).

Figure 4.10 presents the respective DET curves and Table 4.2 the EER, and the optimal values of $D\hat{C}F$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ for the test set. When using cepstral features alone, EER was 3.78% and $D\hat{C}F$ was 3.65%. MaxRel denotes the system based on the first 52 maximal relevance modulation spectra (MRMS) features, which yielded an EER of 4.98% and a $D\hat{C}F$ of 4.88%. Fusion in the last column refers to the concatenation of the augmented MFCC and the 16 MRMS feature vectors ($78 + 16 = 94$ features). Fusion reduced the EER to 3.14% and $D\hat{C}F$ to 2.97% which is an improvement of $\sim 17\%$ and $\sim 19\%$, respectively, over the augmented MFCC.

Performance of speech detection systems on broadcast news audio in other NIST datasets, typically corresponds to a P_{miss} of $\sim 1.5\%$ and a P_{false} of $1\% - 2\%$ [14, 131, 139]. Here, we report a P_{miss} value of $\sim 2.91\%$ and a P_{false} value of $\sim 3.12\%$, which are both higher than the corresponding published values. We believe that this difference is due to the fact that we used just two classes (speech/nonspeech) while in general more classes are considered (speech plus music, speech and noise etc., see references in [131]). The use of more classes will minimize the false rejection of speech (i.e., P_{miss}) when noise or music is also present with speech, because these extra classes could be subsequently reclassified as speech [131]. In addition, several hours of data are commonly used for training of a speech/nonspeech detector [5, 139] whereas we only used about one hour of data.

Table 4.2: *$D\hat{C}F$, $\hat{P}_{miss}$ and $\hat{P}_{false}$ for testing on NIST RT-03*

	MFCCs+$\Delta + \Delta\Delta$	MaxRel	fusion
EER	3.78	4.98	3.14
$D\hat{C}F$	3.65	4.88	2.97
$\hat{P}_{miss}$	3.38	4.62	2.91
$\hat{P}_{false}$	4.40	5.60	3.12

Comparing Tables 4.1, 4.2, we conclude that system performance is better in terms of EER and accuracy in the NIST database than in the Greek broadcast audio data. By inspection of

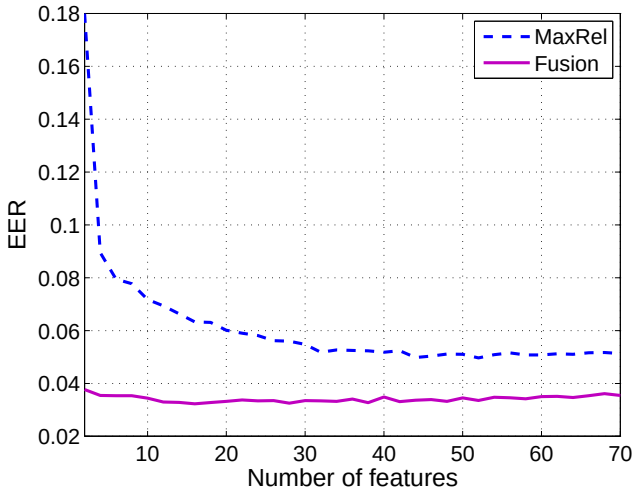


Figure 4.9: SVM classifier equal error rate (EER) as a function of most relevant modulation spectral features alone, or using them in combination with MFCC features for the U.S. English validation dataset.

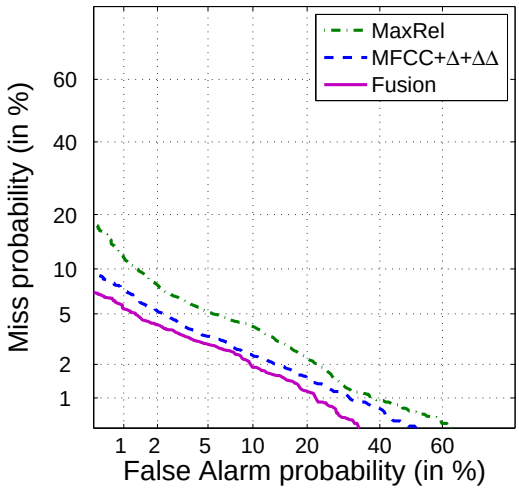


Figure 4.10: DET curves for segment-based SVM classification using the 52 most relevant features (MaxRel), the augmented MFCC features, and Fusion (concatenation of 16 MaxRel with augmented MFCC feature vectors) for the U.S. English test dataset.

the DET curves in Figures 4.8, 4.10, we notice that the lower false alarm regions of the DET curve correspond to higher P_{miss} (false speech rejection) in the Greek dataset than in NIST; on the other hand, P_{false} is lower in the Greek dataset for the lower miss probability regions. This difference in performance could be explained by the different content of the U.S. English and Greek TV shows, i.e., the variability in speech and non-speech classes in every database.

Besides, the concatenation of features yields greater improvement over cepstral features in the NIST database (accuracy $\sim$ 19%, EER $\sim$ 17%) than in Greek broadcast audio data (accuracy $\sim$ 6%, EER $\sim$ 7%).

4.5 Discussion - Conclusion

Previous studies have shown the importance of joint acoustic and modulation frequency concept in signal analysis and synthesis, as well as single-channel talker separation and coding applications ([8, 120, 126]). We presented a dimensionality reduction method for modulation spectral features which could be tailored to various classification tasks. HOSVD efficiently addresses the differing degrees of redundancy in acoustic and modulation frequency subspaces. By projecting features on a lower dimensional subspace, we significantly reduce computational load of MI estimation. Using HOSVD alone would lead to feature selection based minimal redundancy irrespective of their discriminative power [101].

The set of most relevant features exhibited rather comparable classification performance to that of state-of-the-art mel cepstral features (see Figures 4.8& 4.10). Feeding the fused feature set into the same SVM classifier that we used before, further decreased the classification error across the DET curve which supports the hypothesis that modulation spectral features provide non-redundant information to that encoded by MFCCs (Tables 4.1& 4.2).

The suggested features span a segment of 500 ms which is roughly equivalent to two syllables duration; hence, they can capture sound patterns present in a language and that is how they complement MFCC features. On the other hand, this is a non-desirable aspect when we want to use the same system for different languages since further training may be necessary.

Modulation spectra have found important applications to classification tasks such as content identification [126], speaker recognition [64, 108], etc. We have also performed singing voice discrimination from other harmonic musical instruments using Modulation Frequency Features [83]. In that work, a feature set derived from modulation spectra was applied to the task of detecting singing voice in historical and recent recordings of Greek Rembetiko, using an SVM classifier. Modulation features were found to be inferior to MFCCs and their delta coefficients, the feature set most suitable for this task according to a recent comparative study [115]. Fusion of the proposed features with MFCCs and delta coefficients, reduced the optimal detection cost from 11.11% to 9.01% which was rather a modest improvement. Still, the enhanced classification performance in the low false alarm region, could be important in singer recognition applications.

We have also presented an initial evaluation of modulation spectral features for speaker verification and identification in [84]. A theoretical study followed by experimental verification for feature selection in speaker recognition has shown that there is a close connection between classification error probability and mutual information (MI) between speaker identity and features [36]. Hence we addressed the relevance of modulation spectrum to speaker verification and identification using MI [84]. Simulations carried out on YOHO database [17] showed that the relevance of modulation spectral features to this task was rather small and speaker-dependent; a speaker verification system based on modulation spectral features then, should rely on speaker specific features. These might reflect intuitively distinctive features of a speaker such as his/her pitch, a particular manner of speaking, or the nature of glottalization [72]. As the speaker-dependent variability of the mutual information of these features implies, the degree of their significance to speaker recognition and verification and the fusion gain with MFCCs will vary accordingly: it might be minor for some speakers and greater for others with more "atypical" speech. Experimental verification of these results should use other databases than YOHO, with channel mismatch and noise and atypical voices. In the latter case, combination of modulation spectrum with MFCC features might be proven beneficial.

Amplitude-modulation features can capture glottal source differences in normal speech; variation in realization of glottalization of a normal speaker, appears to an extreme degree in dysphonic speech [72]. In the next chapter, we report on detecting dysphonic voices using modulation based features.

Chapter 5

Pathological Voice Quality Assessment

Synopsis of the Chapter

In this Chapter, we explore the information provided by the joint acoustic and modulation frequency representation, referred to as Modulation Spectrum in Chapter 2, for the detection and discrimination of voice disorders. In addition, objective voice quality assessment is addressed in the last section. The initial representation is first transformed to a lower-dimensional domain using higher order singular value decomposition (HOSVD) as described in Chapter 3. From this dimension-reduced representation, a feature selection process is suggested using an information theoretic criterion based on the Mutual Information between voice classes (i.e., normophonic/dysphonic) and features. To evaluate the suggested approach and representation, we conducted cross-validation experiments on databases of sustained vowel recordings from healthy and pathological voices.

This chapter is based upon the following publications:

- Markaki M. and Stylianou Y., *Using Modulation Spectra for Voice Pathology Detection and Classification*, IEEE EMBC, 2009.
- Markaki M. and Stylianou Y., *Normalized Modulation Spectral Features for Cross-Database Voice Pathology Detection*, InterSpeech, 2009.
- Markaki M. and Stylianou Y., *Modulation Spectral Features for Objective Voice Quality Assessment: the Breathiness case*, MAVEBA, 2009.
- Markaki M., Stylianou Y., Arias-Londono J.D. and Godino-Llorente J.I., *Dysphonia Detection based on Modulation Spectral Features and Cepstral Coefficients*, ICASSP, 2010.
- Markaki M. and Stylianou Y., *Modulation Spectral Features for Objective Voice Quality Assessment*, IEEE ISCCSP, 2010.
- Arias-Londono J.D., Godino-Llorente J.I., Markaki M. and Stylianou Y., *On combining information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for automatic detection of pathological voices*, Logopedics Phoniatrics Vocology, 2010.
- Markaki M. and Stylianou Y., *Voice Pathology Detection and Discrimination Based on Modulation Spectral Features*, IEEE Transactions on Speech and Audio Processing, 2011.

I would like to thank Julian Arias-Londono (Universidad Politécnica de Madrid) for his collaboration.

5.1 Introduction

Many studies have focused on identifying acoustic measures that highly correlate with pathological voice qualities (also referred to as voice alterations). Using acoustic analysis, we seek to objectively evaluate the degree of voice alterations in a noninvasive manner. Objective voice quality assessment can assist the perceptual evaluation of dysphonic voice quality used by the clinicians. The most common systems of pathological voice description refers to the degree of "hoarseness" [56]. Hoarseness is perceptually related to the noise generation during phonation. The degree of voice hoarseness can be quantified according to the GRASB (grade, roughness, asthenicity, strain and breathiness) scale proposed by Hirano [56].

The definition of these quantifiable perceptual dimensions (GRASB parameters) is related to a set of descriptive parameters for acoustic phenomena. The perceived voice abnormality is assumed to originate at the vocal source rather than resulting from abnormalities in the vocal tract configuration. Consequently, abnormal vibration patterns and increased turbulent airflow at the level of the glottis might be observed [11]. Acoustic parameters that quantify the glottal noise include fundamental frequency, jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), harmonics to noise ratio (HNR), normalized noise energy (NNE), voice turbulence index (VTI), soft phonation index (SPI), frequency amplitude tremor (FATR), glottal to noise excitation (GNE)([24, 104, 100] and references within).

Some of the suggested features require accurate estimation of the fundamental frequency which is not a trivial task in the case of certain vocal pathologies. Moreover, since these features refer to the glottal activity, an estimation of the glottal airflow signal is required. This can be obtained either by electroglottography (EGG) [42] or by inverse filtering of speech [106] [116] where an estimate of the glottal airflow signal is obtained. Based on the second approach, spectral related features have been defined such as the spectral flatness of the inverse filter (SFF) and the spectral flatness of the residue signal (SFR) [24]. Flatness is defined as the ratio of the geometric mean of the spectrum to its arithmetic mean (usually in dB) [24]. The more noise-like a speech signal is, the larger is the "flatness" of its magnitude spectrum [93]. SFF and SFR can be considered as a measure of the noise masking formants and harmonics, respectively [104].

Apart from the above measurements, there is a great interest in applying methods from the non-linear time series analysis to speech signals, trying to quantify in a compact way the high degree of abnormalities observed during sustained phonation when dysphonia is present. Correlation dimension and second-order dynamical entropy measures [145], Lyapunov exponents [46],

higher-order statistics [3], and measures based on time-delay state-space recurrence and detrended fluctuation analysis [69] have also been used in classifying normophonic from dysphonic speakers. For an extended summary on nonlinear approaches for voice pathology detection, the interested reader is referred to [69].

Assuming that the speech signal production is based on the well-known source-filter theory, then it is expected that perturbations at the glottal level (source signal) will affect the spectral properties of the recorded speech signal. In this case, the estimation of the glottal signal is not necessary. Nevertheless, another difficult problem is raised; the estimation of appropriate features from the speech signal which are connected with properties of the glottal signal. Alternatively both parametric and non-parametric approaches have been suggested in this respect, these being generally referred as *Waveform Perturbation* methods (even if they only work with a partial information of the waveform, i.e., magnitude spectrum, frequency perturbations, etc.). The parametric approaches are based on the source-filter theory for the speech production and on the assumptions made for the glottal signal (i.e., impulse train, noise-like) [27] [6]. The non-parametric approaches are based on the magnitude spectrum of speech where short-term mel frequency cepstral coefficients (MFFC) are widely used in representing the magnitude spectrum in a compact way [29] [28] [48] [47]. The non-parametric approaches also include time-frequency representations as the one suggested in [133].

Correlation of the various suggested features and representations with voice pathology is evaluated using techniques like linear multiple regression analysis [104], or likelihood scores using Gaussian Mixture Models (GMM) [29] [48] and Hidden Markov Models (HMM) [28]. Also neural networks and Support Vector Machine classifiers have been suggested [47] [58].

While there are many suggested features and systems for voice pathology detection in the literature, there have been a few attempts towards separating different kinds of voice pathologies. Linear Prediction-derived measures were found inadequate for making a finer distinction than the normal/pathological voice discrimination in [104]. In [116], after applying an iterative residual signal estimator, features like jitter have been computed. Jitter provided the best classification score between pathologies (54.8% for 21 pathologies). In [28], an HMM approach using MFCC provided an average score of correct classification of 70% (5 pathologies, multi classification experiment).

In [72] a vocal-fold paralysis recognition system using amplitude-modulation and MFCC features combined with GMM, provided an Equal Error Rate (EER) of $\sim 30\%$ in the best case. A recent study for the discrimination of voice pathology signals was carried out using adaptive

growth of Wavelet Packet tree, based on the criterion of Local Discriminant Bases (LDB) [58]. A genetic algorithm was employed to select the best feature set and then a Support Vector Machine (SVM) classifier was used. An average detection score of 83.9% was reported in classifying vocal polyps against adductor spasmodic dysphonia, keratosis leukoplakia, and vocal nodules.

In this chapter, we suggest the use of modulation spectra for detection and classification of voice pathologies [55, 8]. Modulation spectral features have been employed for single-channel speaker separation [120], for speech and speaker recognition [50, 62] as well as for content-based audio identification [126] and speech detection [81]. There are a few works which make use of modulation spectra for voice pathology detection [72] [87], [86]. Modulation spectra may be seen as a non-parametric way to represent the modulations in speech. Modulation spectra offer an implicit way to fuse the various phenomena observed during speech production, such as the harmonic structure during voiced phonation etc. [120]. This is achieved by describing the joint distribution of energy across different acoustic and modulation frequencies. The long-term ($\sim 200 - 300$ms) information that modulation spectrum represents poses a serious challenge to classification algorithms because of its high dimensionality. Past research has addressed the problem of reducing modulation spectral feature dimensions by simple averaging [72], or using modulation scale analysis, a joint representation of the acoustic and modulation frequency with nonuniform bandwidth [126]. In [64], a bank of mel-scale filters has been applied along the acoustic frequency dimension, and discrete cosine transform (DCT) along the modulation frequency axis.

We compute modulation spectra using simple Fourier transform in both frequency axes (acoustic and modulation frequency), as described in Chapter 2. Moreover, we approach the dimensionality reduction of the acoustic and modulation frequency subspaces in the framework of multilinear algebra (cf Chapter 3). Since the acoustic and modulation spectra are characterized by varying degrees of redundancy, we address dimensionality reduction separately in each subspace using higher order singular value decomposition (HOSVD) [68]. The Mutual Information (MI) measurement based on Information Theory [22] can subsequently analyze the relation between the compact lower dimensional features and classes (cf Chapter 3).

Section 5.2 motivates the use of modulation frequency analysis for voice pathology detection and classification, by providing examples of this joint frequency representation computed for speech signals generated by normophonic and dysphonic speakers. For this purpose, speech examples from the Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory (MEEI) database [37] are considered. In Section 5.3, the ability of modulation frequency features to

distinguish between normal and pathological voices is investigated in the first experiment. Next, we investigate the ability of modulation spectra and the suggested feature selection algorithm to make distinctions that are finer than the normal/pathological dichotomy. Specifically we address the binary discrimination between vocal fold polyp, adductor spasmodic dysphonia, keratosis leukoplakia, vocal nodules, as well as between paralysis and all the above voice disorders. Also a general description of MEEI [37] database is provided along with its subsets used in the classification experiments (related papers [87], [89]).

In Section 5.4, we present another set of experiments which tests the performance of the trained dysphonia detector on unknown (completely unseen) data, in the sense that these data are not part of the initial database (related papers [86, 90, 4]). Unseen data may have been recorded under different conditions from those of the initial database which was used for training. For this purpose, we used a second database provided to us by Universidad Politécnica de Madrid, which is referred to as Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [49]. Similar to MEEI, PdA contains recordings of sustained vowels /a/ and was developed for voice function assessment purposes.

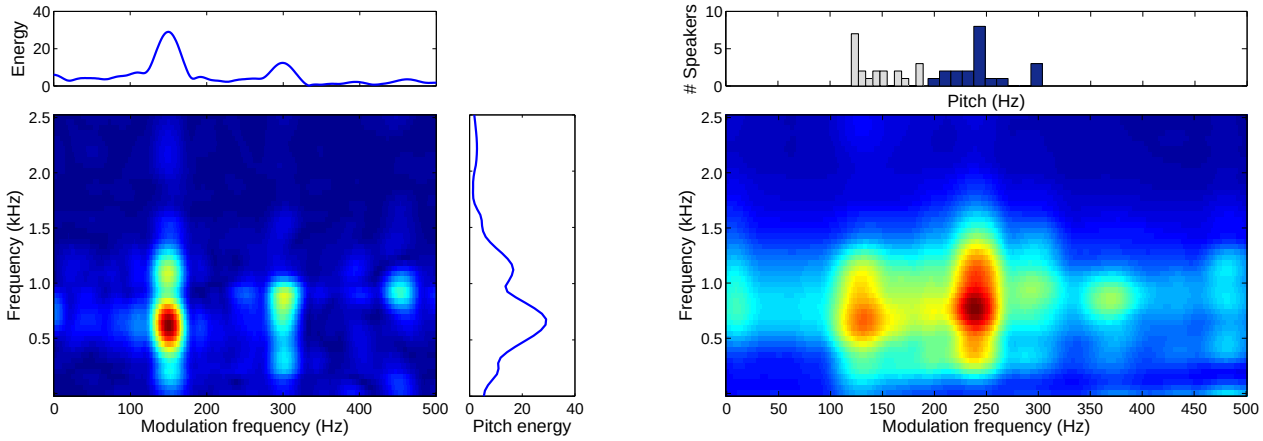
In Section 5.5, we investigate the correlation of modulation spectral features to the degree of grade or hoarseness, (G), of pathological voices. G is the overall degree of deviance of voice, the amount of noise in the produced sound. We conducted experiments on hoarseness classification of data from PdA, using a combination of two naive bayes (NB) classifiers based on different feature sets (related papers [85, 88]).

Finally, conclusions are drawn and future directions are indicated in Section 5.6.

5.2 Modulation Spectral Patterns in Normal and Dysphonic Voices

We have evaluated features of the modulation spectrogram of sustained vowel /AH/ for voice pathology detection and classification tasks. As explained in the work of Vieira et al [134], sustained vowel phonations at comfortable levels of fundamental frequency and loudness are useful from a clinical point of view. In addition, the time domain acoustic signal of /AH/ exhibits larger and sharper peaks than the other vowels; these signal features are well correlated to the electroglottal graph (EGG) parameters.

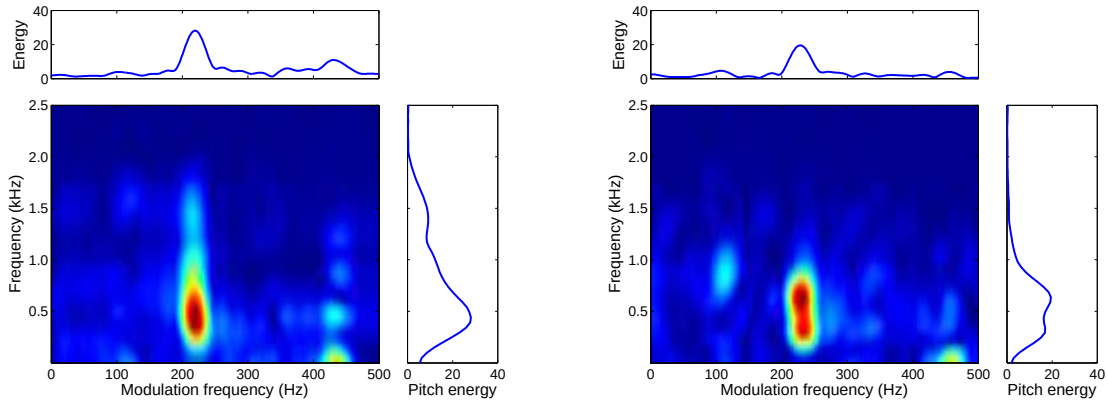
Fig. 5.1a shows the modulation spectrogram $|X_l(k,i)|$ of a 262 ms long frame from sustained phonation speech samples of the vowel /AH/ uttered by a normal male speaker from the MEEI database [37]. Apparently these phonations do not possess the syllabic and phonetic temporal



(a) (b)

Figure 5.1: (a) Modulation spectrogram of sustained vowel /AH/ by a 34 years old normal male speaker ($\sim$ 150 Hz fundamental frequency). The two side plots present the slices intersecting at the point of maximum energy; its coordinates coincide with the fundamental frequency and the first formant of /AH/ ($\sim$ 590 Hz). Vertical plot displays the localization of fundamental frequency energy at vowel formants along the acoustic frequency axis; the upper horizontal plot presents the energy localization of first formant at the fundamental frequency and its harmonics along the modulation frequency axis. (b) Mean values for the modulation spectra of 40 normal speakers from MEEI database [37]. The number of male equals the number of female subjects. All modulation spectra have been normalized to 1 prior to averaging. Upper horizontal plot displays the histogram of fundamental frequency values of male (grey) and female normal speakers (black).

structure of speech. Hence, the higher energy values are not concentrated at the lower modulation frequencies which are typical in running speech, $\sim 1-20$ Hz [50]. Instead, since we used an analysis window $h(n)$ that was shorter than the expected lowest pitch period, the highest energy terms usually occur at the fundamental frequency of the speaker ($\sim$ 150 Hz in the example shown in Fig. 5.1a) and its harmonics in the modulation frequency axis (up to 500 Hz). Fundamental frequency energy appears localized at the first two formants of vowel /AH/ along the acoustic frequency axis (their range is $\sim 677 \pm 95$ Hz and $\sim 1083 \pm 118$ Hz). Fig. 5.1b displays the mean modulation spectrum, and fundamental frequency distribution of 40 normal speakers from MEEI, with equal number of male and female subjects. All modulation spectra have been normalized to 1 prior to averaging. The two main clusters reflect the fundamental frequency distribution of male (range: 146 ± 24.4 Hz) and female talkers (244 ± 30 Hz). The second cluster contains more energy than the first cluster, since it also comprises energy from the first harmonic of the fundamental frequency of male speakers. Regarding the vertical coordinates of clusters, most energy is concentrated around the first two formants of /AH/. Overall, modulation spectral



(a) Vocal Polyps

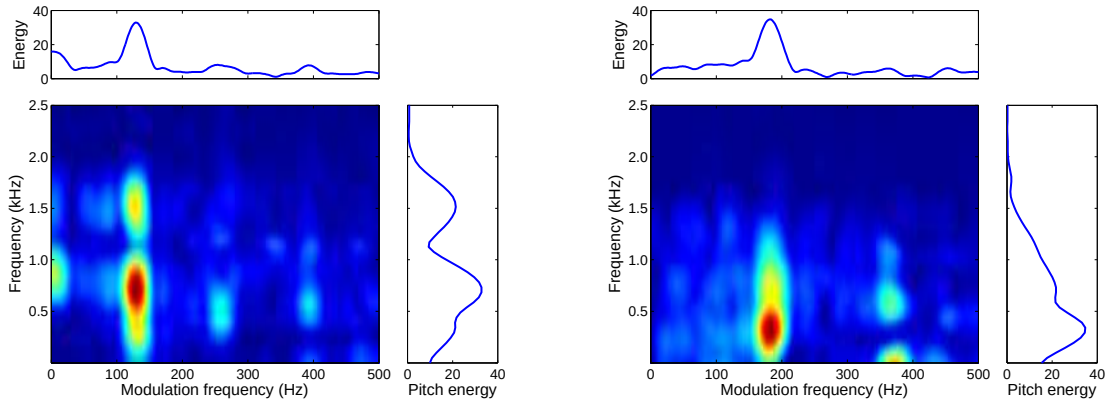
(b) Adductor spasmodic dysphonia

Figure 5.2: Modulation spectrogram of (a) a 39 years old woman with vocal polyps ($\sim$ 220 Hz fundamental frequency), (b) a 49 years old woman with adductor spasmodic dysphonia ($\sim$ 230 Hz fundamental frequency).

representations of normal vowel phonations are quite similar to each other, exhibiting a clear harmonic structure.

These patterns of amplitude modulations are expected to be distorted when voice pathology is present - providing therefore cues for its detection and classification. Fig. 5.2 and 5.3 depict modulation spectra $|X_l(k,i)|$ of sustained vowels produced by patients with various voice pathologies: vocal polyps, adductor spasmodic dysphonia, keratosis and vocal nodules. A comprehensive description of these pathologies is provided in [11]. Polyps are solid or fluid filled growths arising from the vocal fold mucosa. They affect vibration of vocal folds depending on their size and location. In adductor spasmodic dysphonia vocal folds suddenly squeeze together very tightly and in effect the voice breaks, stops, or strangles. Keratosis refers to a lesion on the mucosa of the vocal folds, appearing as a white patch. Nodules are swellings below the epithelium of vocal folds; they might prevent the vibration of the vocal folds either by causing a gap between the two vocal folds - which lets air to escape - or by stiffening the mucosal tissue.

Compared to the normal ones (see Fig. 5.1), pathological modulation spectra lack a uniform harmonic structure and appear more "spread" and "flattened" across the acoustic frequency axis. Main differences can be spotted near the low acoustic frequency bands where the first formant of /AH/ is located ($\sim$ 500 Hz). In the polyp case (Fig. 5.2a), the maximum energy is located below the first formant in the acoustic frequency axis, close to its fundamental frequency in the modulation frequency axis ($\sim$ 220 Hz). In the case of the speaker with adductor spasmodic dysphonia, we also observe the strong modulations of the first formant by the fundamental



(a) Keratosis (b) Vocal Nodules

Figure 5.3: Modulation spectrogram of (a) a 50 years old female speaker with keratosis leukoplakia ($\sim$ 135 Hz fundamental frequency). (b) a 38 years old female speaker with vocal nodules ($\sim$ 185 Hz fundamental frequency).

frequency (230 Hz) of the speaker. However, in this case, there is important energy in a frequency lower than the 1st formant (280 Hz) which is also modulated by the fundamental frequency. For this speaker, there are strong subharmonics. Fig. 5.2b shows then that there are noticeable modulations (although not as strong as for the fundamental frequency) of the 2nd formant (900 Hz) by these subharmonics (115 Hz) (see Fig. 5.2b). Some differences are also observed at larger modulation frequencies, which correspond to the harmonics of these fundamental frequency values (Fig. 5.2a, 5.2b and 5.3b). High energy might appear at modulations lower than $\sim$ 30 Hz, near the first formant as in the case of keratosis (Fig. 5.3a); there is also high energy beyond the second formant ($\sim$ 1100 Hz) located near the fundamental frequency value in the modulation axis ($\sim$ 134 Hz).

In short, the high resolution of modulation spectral representation yields quite distinctive patterns depending on the type and the severity of voice pathology allowing thus a finer than normal/abnormal distinction.

5.3 Experiments on Dysphonia Detection and Classification

5.3.1 Database

The database we used was designed to support the evaluation of voice pathology assessment systems; it was developed by Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory

Table 5.1: Normal and pathological talkers [100]

	Number		Mean age (years)		Standard deviation (years)	
Talkers	Male	Female	Male	Female	Male	Female
Normal	21	32	38.8	34.2	8.5	7.9
Pathological	70	103	41.8	37.4	9.3	8.1

and it is referred to as MEEI database [37]. The database contains sustained vowel samples of ~ 3 sec duration from 53 normal talkers and of ~ 1 sec duration from 657 pathological talkers with a wide variety of organic, neurological, traumatic, psychogenic and other voice disorders. The database also includes voice samples of ~ 12 sec duration of the same subjects reading text from "Rainbow passage" and it is commercially available.

For the first test case, we used the sustained vowel phonations from a subset of MEEI, referred to as MEEI$_{sub}$, first defined in [100]. MEEI$_{sub}$ includes 53 normal and 173 pathological speakers with similar age and sex distributions avoiding therefore any bias by these two factors. Pathological class includes many different voice disorders. Since the ratio of the normal to pathological talkers in MEEI$_{sub}$ (~ 0.3) is quite close to the inverse ratio of the respective vowel durations, the number of segments in each class is close enough: 2240 samples of normal voices, vs 1864 samples of pathological ones. Statistics of this subset of MEEI database are provided in Table 5.1.

For voice disorder discrimination, two different kinds of experiments were performed. The first series of experiments consisted of discrimination between a pair of different pathologies. For comparison purposes, the same subset of pathologies as the one considered in [58] was selected: vocal fold polyp, adductor spasmodic dysphonia, keratosis leukoplakia, and vocal nodules. A full pairwise classification was performed as opposed to [58] where only the binary discrimination of vocal fold polyp against the three other pathologies has been reported. There were 88 such cases in the whole MEEI database; only 49 out of these speakers were included in MEEI$_{sub}$ dataset. There was a co-occurence of two pathologies at the same person in 5 cases, making a total of 83 subjects The last experiment consisted of the discrimination of vocal fold paralysis from all the above mentioned pathologies. There were 71 paralysis cases in MEEI with no co-occurence of the other four disorders (refer to Table 5.2 for statistics). These were compared to 71 cases characterized by at least one of the four disorders.

Most of the selected recordings had a sampling rate of 25 kHz; files with a 50 kHz sampling rate were antialias-filtered and downsampled to 25 kHz. Each file was partitioned into 262 ms

Table 5.2: Number and Sex of Patients Included in Medical Diagnosis Categories

Medical diagnosis	No. of males	No. of females	No. of segments
Vocal Nodules	1	19	212
Vocal Polyp	12	8	220
Keratosis	15	11	253
Adductor	3	19	232
Paralysis	35	36	781

segments for long-term feature analysis; evenly spaced overlapping segments were extracted every 64 ms similar to [120]. This frame rate can capture the time variation of amplitude modulation patterns evident in each frequency band.

5.3.2 Methods

Modulation spectra were computed using the Modulation Toolbox [7] throughout all experiments. Wideband modulation frequency analysis was considered so that an adult speaker's fundamental frequency could be resolved in the modulation frequency axis [120]. Hence, the variables in eq. (2.13) and (2.13) in Chapter 2, were set as following: $M = 25$ samples (1 ms time-shift at 25 kHz sampling frequency), $L = 38$ samples, $K = 512$, and $I = 512$; $h(n)$ and $g(m)$ were a 75-point (or, 3 ms) and 78-point Hamming window, respectively. The equations are repeated below for the sake of completeness:

$$\begin{aligned} X_k(m) &= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_K^{kn}, \\ k &= 0,\ldots,K-1, \end{aligned} \tag{5.1}$$

where $W_K = e^{-j(2\pi/K)}$, $h(n)$ the (acoustic) frequency analysis window and M the hopsize (in number of samples), and

$$\begin{aligned} X_l(k,i) &= \sum_{m=-\infty}^{\infty} g(lL-m)|X_k(m)|W_I^{im}, \\ i &= 0,\ldots,I-1, \end{aligned} \tag{5.2}$$

where $W_I = e^{-j(2\pi/I)}$, $g(m)$ is the modulation frequency analysis window and L the corresponding hopsize (in number of samples); k and i are referred to as the "Fourier" (or acoustic) and "modulation" frequency, respectively.

One uniform modulation frequency vector was produced in each one of the 257 subbands.

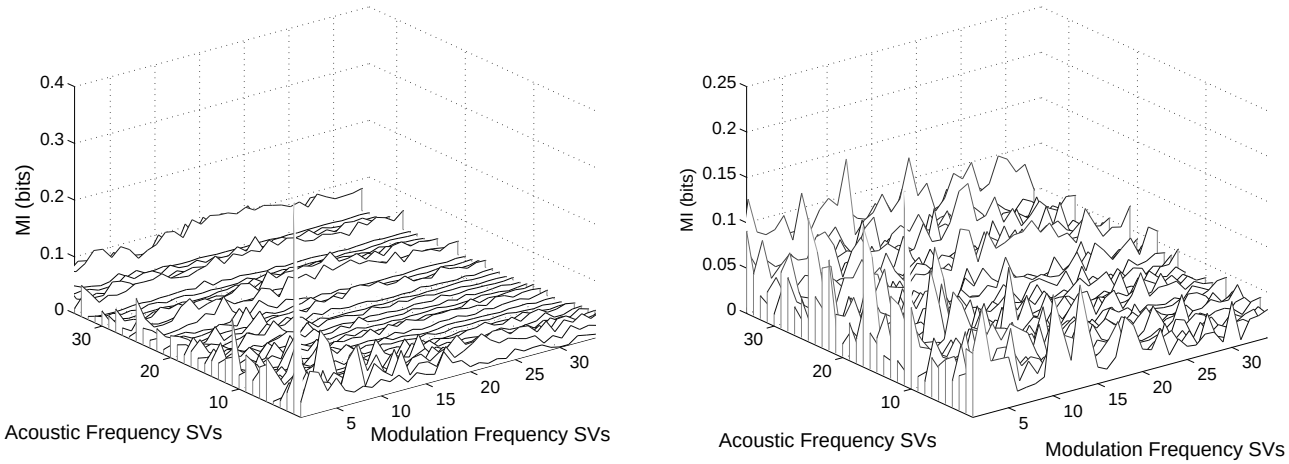
Due to the 1 ms time-shift (window shift $M = 25$ samples) each modulation frequency vector consisted of 257 (up to π) elements up to 500 Hz.

For the computation of the singular matrices for HOSVD, a random subset of 25 normophonic and 25 dysphonic speakers was selected once. Using 1s from each speaker, and considering segments of 262ms for the computation of modulation spectra, with a shift of 64ms, 12 modulation spectra matrices of dimension 257×257 each, were generated per speaker. Stacking the $12 \times 50 = 600$ modulation spectra matrices for all the speakers in the above subset, produced the data tensor $\mathcal{D} \in \mathbb{R}^{257 \times 257 \times 600}$. Before applying HOSVD, the mean value of the tensor was computed and then subtracted from the tensor.

The singular matrices $\mathbf{U}^{(1)} \equiv \mathbf{U}_{af} \in \mathbb{R}^{257 \times 257}$ and $\mathbf{U}^{(2)} \equiv \mathbf{U}_{mf} \in \mathbb{R}^{257 \times 257}$ were directly obtained by SVD of the "matrix unfoldings" $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$ of $\mathcal{D}$ respectively. The singular vectors which exceeded a contribution threshold of 0.2% were retained in each mode (eq. 3.4), resulting in the truncated singular matrices $\hat{U}_{af} \in \mathbb{R}^{257 \times 34}$ and $\hat{U}_{mf} \in \mathbb{R}^{257 \times 34}$. It is worth noting that the above process to compute the truncated singular matrices using HOSVD was performed only once. HOSVD is the most costly process in our system since it consists of the SVD of the two data matrices $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$, with dimension $N \times k$ each. Note that the computational complexity of SVD transform is $O(Nk^2)$. N is either the acoustic frequency dimension or the modulation frequency dimension; respectively, k is the product of the modulation or the acoustic frequency dimension multiplied by the size of the training dataset ((i.e., $k = 257 \times 600$ in this case). The truncated matrices were saved and used for all the detection and classification experiments. Features were projected on $\hat{U}_{af}$ and $\hat{U}_{mf}$ according to eq. (3.5) resulting in matrices $\mathbf{Z} \in \mathbb{R}^{34 \times 34}$; these were subsequently reshaped into vectors before MI estimation, feature selection, and SVM classification.

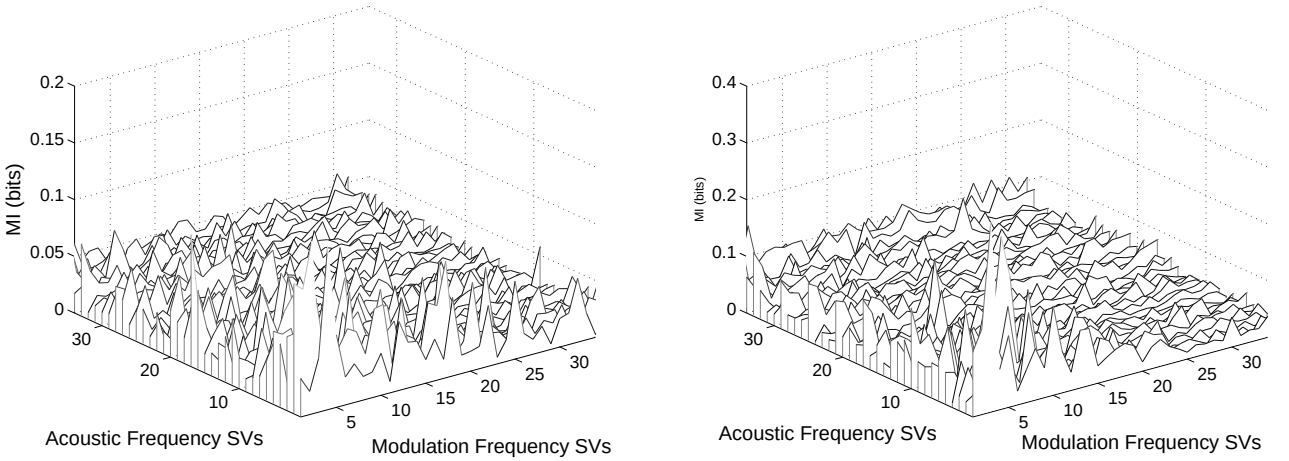
For the data discretization involved in MI estimation, the number of discrete bins along each axis was set to $b^* = 8$ according to the procedure described in [123]. Through a sequential search, the top m features in the descent ordering of $I(x_j; c)$ - i.e., the most relevant features - were selected in every case [101]. We computed the cross-validation classification error (EER) for an increasing number of these sequential features in order to determine the optimal size of feature set, m.

Fig. 5.4 and 5.5, present the MI estimates between reduced features and the class variable in the four (out of 8) different classification tasks. In the normal vs pathological case and the polyp vs nodules case, the MI of the most relevant features is ~ 0.35 and ~ 0.3 bits, respectively, and the number of relevant features is small. For polyp/adductor discrimination MI is ~ 0.2 bits whereas



(a) (b)

Figure 5.4: Mutual information (MI) values (a) for the normal vs pathological voice classification task; (b) for the polyp vs adductor classification task.



(a) (b)

Figure 5.5: Mutual information (MI) values (a) for the polyp vs keratosis classification task; (b) for the polyp vs nodules classification task.

for polyp/keratosis discrimination MI is ~ 0.14 bits. For adductor/nodules, adductor/keratwsis and keratwsis/nodules discrimination, the corresponding values of MI are ~ 0.18, ~ 0.25 and ~ 0.28 bits, respectively. However, the MI is significantly lower for the discrimination of paralysis against the other four disorders: its maximum value is only ~ 0.06 bits. This is due to the fact that the non-paralysis signals include several other disorders (four at least) so there is not an

homogeneity in the non-paralysis class. Hence, it is very difficult in this case to find optimum features in terms of relevance as in the other binary classification cases. The absolute scale of MI is actually a predictor of the performance of the classification system based on the maximum relevance feature selection scheme as it will be shown next [22].

5.3.3 Results

Eight binary classification tasks were defined that exploit the patterns of energy distribution in modulation spectra: normal vs abnormal phonation, a full pairwise comparison between four voice disorders (vocal polyps, adductor spasmodic dysphonia, keratosis, vocal nodules), and paralysis vs the combined previous four disorders.

Classification performance was computed when vector components were selected based on maximum contribution (maxContrib) (eq.3.4), or maximum relevance (maxRel) criteria. Pattern classification was carried out using Support Vector Machine (SVM) classifiers. SVM find the optimal boundary that separates two classes maximizing the margin between separating boundary and closest samples to it (support vectors) [60]. In this work, SVMlight [60] with a Radial-Basis-Functions kernel was used. Tests with linear SVM with or without spherical normalization were also conducted. This is a modified stereographic projection recommended before classification of high dimensional vectors using linear SVM [136].

A 4-fold stratified cross-validation was used, which was repeated 40 times. The classifier was trained on the 75% of speakers of both classes, then tested using the remaining 25%. MI estimation using (randomly chosen) 75% of each dataset during 4-fold stratified cross-validation gives almost identical results with MI estimation based on the full dataset. Training and testing was based on 262ms segments; utterance classification was then computed using the median of the decisions over its segments.

The system performance was evaluated using the detection error trade-off curve (DET) between false rejection rate (or miss probability) and false acceptance rate (or false alarm probability) [94]. The rates of each type of errors depend upon the value of a threshold, T. The optimal detection accuracy (DCF_{opt}) occurs when T is set such that the total number of errors is minimized. DCF_{opt} reflects performance at a single operating point on the detection error trade-off (DET) curve. The Equal Error Rate (EER) refers to the point at the DET curve where the false-alarm probability equals the miss probability. DET curves present more accurately than Receiver Operating Characteristic (ROC) curves the performance of the different assessment systems at the low error operating points [94]. We depict representative DET curves, and report

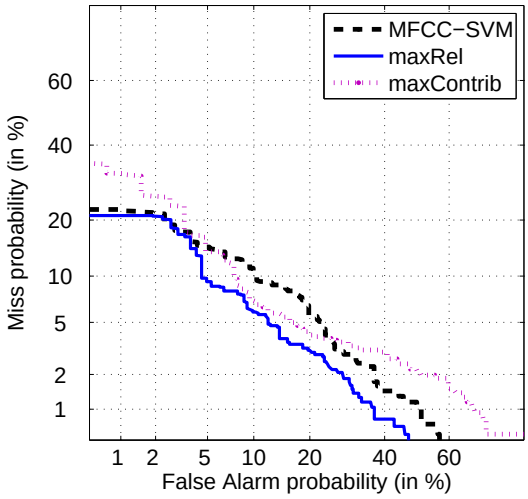


Figure 5.6: DET curves for the dysphonia detection system using $[7 \times 13]$ dimensions according to maximum contribution criterion (red dashed), the system based on the 20 most relevant features (blue solid) and MFCC features (black dotted) with the same SVM classifier.

on DCF_{opt}, EER, and area under the ROC curve (AUC) for the classification tasks, along with their corresponding 95% confidence intervals. Please note that the curves and measures refer to the average of the 40 runs.

In Table 5.3, we present AUC, DCF_{opt}, and EER for the dysphonia detection task, both for segments and utterances along with their corresponding 95% confidence intervals. For the cases of maximum relevance (maxRel) and maximum contribution criterion (maxContrib), the optimum number of features is also provided in parenthesis. For comparison purposes, we present the performance of another system obtained for utterances on the *same data* based on short term mel-cepstral parameters (defined as in [48]) and the same SVM classifier (denoted as MFCC-SVM in Table 5.3). We also present the AUC and the DCF_{opt} of the system described in Godino et al. [48] based on Gaussian Mixture Models (GMM) and MFCC parameters using approximately the same subset of MEEI (this is denoted as MFCC-GMM in Table 5.3). Although the results reported in [48] are better in terms of AUC, the authors have used a somewhat different cross-validation procedure and have kept 147 pathological signals out of the 173 ones which are included in the MEEI subset used in this work [100].

The best system that was based on maxRel used 20 features whereas the best system based on maxContrib used $[7 \times 13] = 91$ features. In Fig. 5.6, we compare the performance of the systems using the same SVM classifier in terms of DET curves. The system that has been built on most relevant features is a little superior compared to the other systems, especially in the lower false

Table 5.3: Area Under the ROC curve (AUC), Efficiency (DCF_{opt}) and Equal Error Rate (EER) for discrimination of Normal and Pathological Talkers using modulation spectra and MFCC features with the same SVM classifier (95% confidence intervals). The last row in the table refers to the corresponding AUC and DCF_{opt} for the same task using MFCC features and GMM as reported in [48].

	Segment (262ms)			Utterance		
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)
max Relevance (20)	0.9656±0.0032	90.43 ± 0.15	8.63±0.57	0.9775±0.0028	94.08 ± 0.28	6.29±0.67
max Contribution [7 × 13]	0.9544±0.0036	89.70 ± 0.23	9.18±0.41	0.9633±0.0035	92.67 ± 0.08	7.50±0.28
MFCC-SVM (40)	0.9626±0.0032	89.60±0.41	10.01±0.54	0.9666±0.0029	91.48 ± 0.37	8.47±0.57
MFCC-GMM [48]	-	-	-	0.9997	94.07 ± 1.05	-

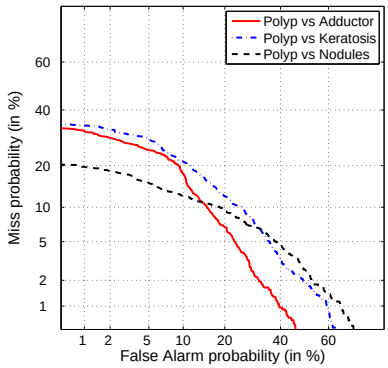


Figure 5.7: DET curves with 4-fold cross-validation using modulation spectral features and SVMs for discrimination between polyp/adductor, polyp/keratosis and polyp/nodules cases in MEEI.

alarm or miss probability regions.

Similar to normal vs pathological discrimination, for the pathology discrimination task the features were first reduced by projecting them on the singular vectors extracted from the same normal and pathological subjects referred to in Table 5.1. The idea was to improve the generalization ability of our pathology classification system. There were less training vectors during the 4-fold cross-validation in all classification tasks. We also tested both strategies for choosing the suitable levels of detail of this representation: maximum contribution and maximum relevance.

Different kernels and spherical normalization [136] yielded marginal differences in classification performance: in general, results were better using RBF kernel than linear kernel. Spherical normalization enhanced results for linear SVMs and large number of features, but this trend was not observed for RBF kernel.

Tables 5.4, 5.5, 5.6 provide the classification per pathology scores in terms of AUC, DCF_{opt}

and EER and the corresponding 95% confidence intervals. For simplicity, only the scores per utterance (or per speaker) are provided. The optimum number of features as this is selected using the maximum relevance or maximum contribution criterion is also presented. For comparison purposes, we report the best discrimination rates (DR) obtained on the *same data* for three classification tasks by Hosseini et al. [58] using SVM on Fisher distance and Genetic Algorithms for feature selection in Table 5.4 (it is denoted as FD-GA). Tables 5.5, 5.6 also present the classification performance of systems based on the standard MFCC features and the same SVM classifier for the other four voice pathology discrimination tasks. Fig. 5.7 presents the DET curves of the system based on most relevant modulation spectral features and SVM for three binary pathology classification tasks.

In every pathology discrimination task, the modulation spectral features were superior to MFCC (see Tables 5.5, 5.6; the results using MFCC for the tasks in Table 5.4 were not included because of lack of space). Except for the paralysis/non-paralysis case (see Table 5.6), classification performance was better when we used most relevant (maxRel) features than features with greatest eigenvalue contribution (maxContribution).

As it was noticed before, the absolute scale of MI could almost "predict" the classification performance of the system based on the maximum relevance feature selection scheme [22]. The MI was significantly lower for the discrimination of paralysis against the other four disorders: its maximum value was only ~ 0.06 bits. There is a trade-off between features relevance and features redundancy in each feature selection technique [101]. When the relevance of individual features towards a classification task is very low then, the minimal redundancy (or, "maximal contribution") criterion obviously prevails. The best EER in the paralysis / non-paralysis discrimination task was $15.45 \pm 0.56\%$ using the $[8 \times 15$ components with maximum contribution vs $27.99 \pm 0.81\%$ (95% confidence intervals) using the 200 most relevant modulation spectral features (Table 5.6). For comparison, the authors in [72] reported an EER of $\sim 30\%$ for the discrimination of paralysis from other voice disorders in MEEI (binary task) based on amplitude modulation features.

5.3.4 Discussion

We have evaluated features of the modulation spectrogram of sustained vowel /AH/ for voice pathology detection and classification. Our results show that modulation spectral features are well suited to voice pathology assessment and discrimination tasks.

In order to extract a compact set of features out of this multidimensional representation, we first removed "redundancy" at the first step of our processing, using HOSVD. HOSVD was

Table 5.4: Area Under the ROC curve (AUC), Efficiency (DCF_{opt}) and Equal Error Rate (EER) per disorder using modulation spectral features and SVM (95% confidence intervals). The corresponding best discrimination rates for the same tasks using FD-GA [58] are listed in the last column of the table.

	max Relevance			max Contribution			FD-GA
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)	DCF_{opt}
Polyp/Adductor	0.9585±0.0087	91.23±1.10 (60)	8.25±1.3792	0.9309±0.0084	81.23±2.11 [17 × 24]	13.25±1.32	82.5%
Polyp/Keratosis	0.9359±0.0058	84.77±1.42 (80)	15.71±1.0848	0.6279±0.0204	57.26±0.25 [17 × 24]	40.25±1.85	81.8%
Polyp/Nodules	0.9428±0.0073	91.66 ± 1.14 (20)	11.25±1.2064	0.8802±0.0127	86.03 ± 1.50 [6 × 10]	16.44±1.29	87.5%

Table 5.5: Area Under the ROC curve (AUC), Efficiency (DCF_{opt}) and Equal Error Rate (EER) for discrimination of different kind of dysphonias using modulation spectral features and MFCC features with the same SVM classifier (95% confidence intervals).

	Adductor / Nodules			Adductor / Keratosis		
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)
max Relevance	0.9578±0.0064	92.09±0.92 (95)	8.44±1.09	0.9949±0.0017	95.77±0.92 (90)	2.17±0.62
max Contribution	0.7981±0.0147	75.21±1.53 12x19	25.46±1.30	0.8844±0.0113	72.15±1.19 12x20	17.21±1.51
MFCC	0.6728±0.0147	63.91±1.03 (20)	37.12±1.32	0.7188±0.0123	66.65±1.81 (20)	36.70±1.47

Table 5.6: Area Under the ROC curve (AUC), Efficiency (DCF_{opt}) and Equal Error Rate (EER) for discrimination of different kind of dysphonias using modulation spectral features and MFCC features with the same SVM classifier (95% confidence intervals).

	Keratosis / Nodules			Paralysis / Other		
	AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	EER (%)
max Relevance	0.9527±0.0053	89.11± 1.33 (97)	13.21±1.18	0.7648±0.0078	70.09±1.05 (200)	27.99±0.81
max Contribution	0.9265±0.0106	86.59± 0.45 [12 × 20]	15.23±1.57	0.9063±0.0052	82.14±0.85 [8 × 15]	15.45±0.56
MFCC	0.7286±0.0183	67.88±0.91 (20)	31.39±1.77	0.6504±0.0081	64.02±0.62 (60)	38.68±0.65

performed on the same dataset of normal and pathological talkers for all classification tasks. The efficiency scores for pathologies discrimination would be better if we had performed HOSVD on pathological samples only. Still we wanted to build a system that could proceed from normal vs pathological discrimination to voice disorder classification, based on features projected on the same principal axes. Features relevance to each task was assessed based on MI estimation.

Classification experiments with MEEI database [37] confirmed that the absolute scale of MI can indeed "predict" the performance of the system based on the maximum relevance feature selection scheme [22]. There is a trade-off between features relevance and features redundancy

in each feature selection technique [101]. When the relevance of individual features towards a classification task is very low then, the minimal redundancy (or, "maximal contribution") criterion obviously prevails. Hence in the last classification task (paralysis/non-paralysis), the maximum contribution features outperformed the maximum relevance features.

5.4 Experiments on different databases

Testing the optimal detector which was developed on MEEI on recordings from a different database (PdA [49] described later in this section), we found that the classification performance was significantly decreased. Apparently, this degradation was caused by the difference of the environmental characteristics - channel transmission effects, noises, etc. - of the training and testing data. Past research has addressed the sensitivity of features to data mismatch with feature normalization [107]. Feature normalization scales or warps the components of the fixed feature vector in order to make both training and testing features independent of environmental characteristics. As a first step towards a robust voice pathology detector, we implement subband normalization of modulation spectral features (cf Chapter 2) that makes them insensitive to time and frequency distortions according to [126]. We validate our approach on normalized modulation spectral analysis through cross-database experiments. Further we investigate the complementary information that *normalized* modulation spectral features provide to mel-frequency cepstral coefficients (MFCC) for voice pathology detection task.

5.4.1 Database

The first dysphonic voice corpus was the Kay Voice Disorders Database [37], referred to as MEEI, which we used in the first set of experiments too.A subset of 173 pathological and 53 normal speakers were selected again according to [133], with similar age and sex distributions. The second database was recorded by Universidad Politécnica de Madrid, and it is referred to as Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [49]. Similar to MEEI, PdA contains recordings of sustained vowels (/a/) and was developed for voice function assessment purposes. For the following experiments, the voices of 200 dysphonic subjects (74 men and 126 women, aged 11 to 76) affected by nodules, polyps, oedema, etc, as well as 199 normal subjects (87 men and 112 women, aged 16 to 70) were used. All the tests were conducted on signals sampled at 25 kHz. A 4-fold stratified cross-validation scheme - repeated 4 times - produced 16 different groupings of the voices, each using $\sim 75\%$ of the utterances for training and

$\sim 25\%$ for testing. For the cross-database evaluation, we used PdA for training and MEEI for testing or vice-versa (in order to simulate the situation of completed unseen, to the classification system, data).

5.4.2 Methods

Normalized modulation spectra

The distribution of envelope amplitudes of voiced speech has a strong exponential component. Hence we calculate modulation spectra using a log transformation of the amplitude values $|X_k(m)|$ and subtracting their mean log amplitude before windowing in (5.3):

$$\hat{X}_k(m) \quad = \quad \log|X_k(m)| - \overline{\log|X_k(m)|} \tag{5.3}$$

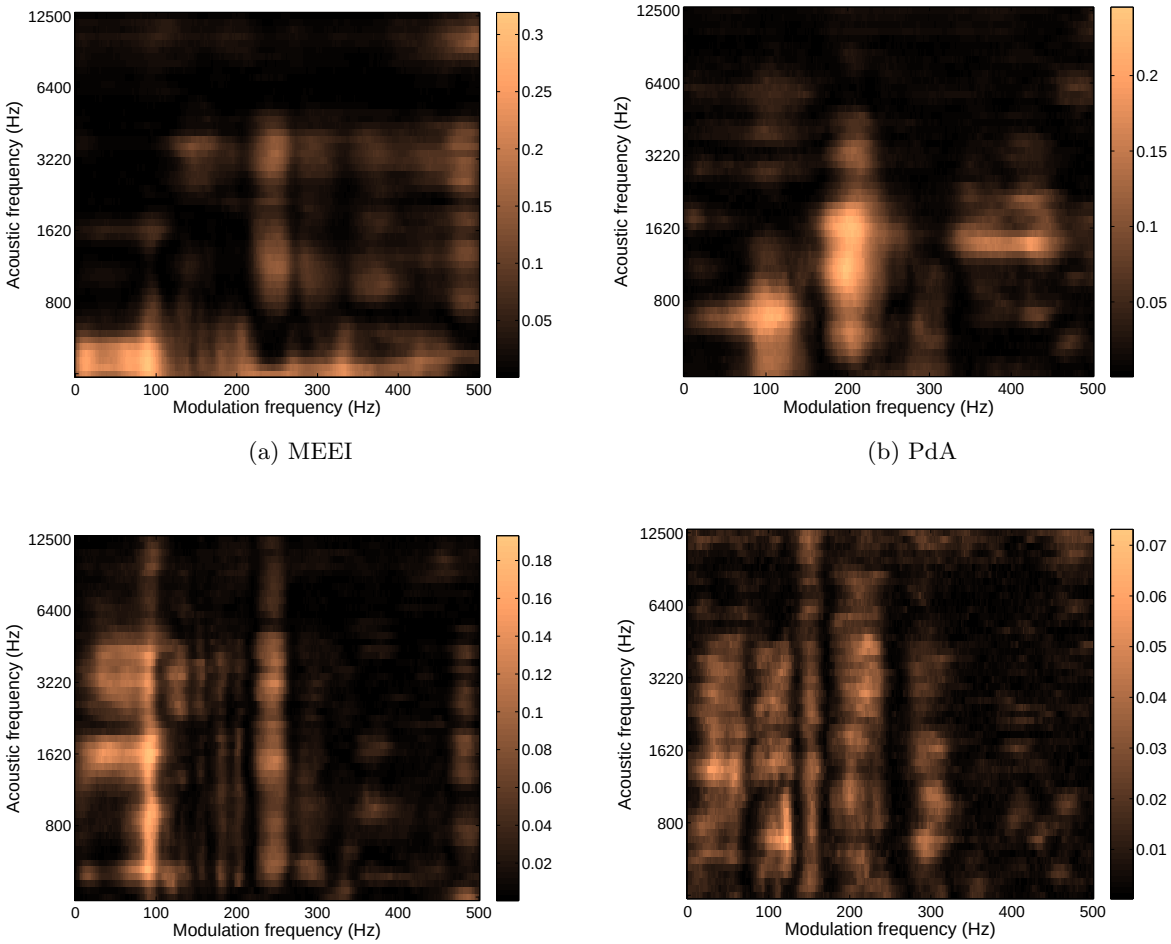
where $\overline{\log|X_k(m)|}$ denotes the average value of $\log|X_k(m)|$ over m. This is analogous to the cepstral mean subtraction approach, which is commonly employed to compensate for convolutional noise in the case of MFCC features. Next, we normalize every acoustic frequency subband with the marginal of the modulation frequency representation:

$$X_{l,sub}(k,i) \quad = \quad \frac{X_l(k,i)}{\sum_i X_l(k,i)} \tag{5.4}$$

As shown in the following, this subband normalization is important when there is a mismatch between training and testing conditions, or in other words, when the detection system is employed in real (testing) conditions.

As a first test of the normalization effects on features sensitivity, we assess the relevance of features to voice pathology detection in MEEI and PdA databases, before and after normalization. As previously, relevance is defined as the mutual information (MI) $I(x_j;c)$ between feature x_j and class c (cf Chapter 3). In general, MI between two random variables x_i and x_j is defined as the KL-divergence between their joint probability density functions (pdf) $P_{ij}(x_i,x_j)$ and the marginal pdf $P_i(x_i)$ and $P_j(x_j)$ [22]. In the case of modulation spectrum representation, the distribution of the MI for a set of features and a given class can be visualized as a picture. In Fig. 5.8a and Fig. 5.8b, the distribution of MI between the selected features and pathologic voices class is depicted, for the MEEI and the PdA databases, respectively, *before* normalization of features. It is then obvious from these two sub-figures that the two distributions of MI are quite different. This means that training a detector on one database and test it on the other database, will result in a

very poor detection performance. Fig. 5.8c and Fig. 5.8d depict the corresponding distribution of MI for both databases *after* feature normalization (Fig. 5.8c for MEEI and Fig. 5.8d for PdA). We observed that after applying the suggested feature normalization the maximum value of MI per database lowers almost by half. This will lead to lower performance detector for each database. However, and compared to the upper panels of the same figure, we observe that the distribution of MI in MEEI is quite comparable to the one obtained in PdA. This means that a robust to unseen data detector is now possible to develop.



(a) MEEI

(b) PdA

(c) MEEI

(d) PdA

Figure 5.8: Relevance (MI) between modulation spectral features and pathologic voice class *without normalization* (a) in MEEI, and (b) in PdA and *after normalization* in (c) in MEEI, and (d) in PdA.

5.4.3 Results

In the following experiments, modulation spectra were computed in a frame-by-frame basis using long windows in time (250 ms) which were shifted by 50ms. We used Mel scale filtering with 53 bands while the size of the Fourier transform for the time-domain transformation was set to 257 (up to π). Therefore, each modulation spectrum consisted of $I_1 = 53$ acoustic frequencies and $I_2 = 257$ modulation frequencies, resulting therefore in an 53×257 image per frame. The normalized modulation spectra computed in each frame were stacked to produce a third order tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the number of frames in the training dataset. Applying the High Order SVD algorithm described previously, the near-optimal projections or principal axes (PCs) of features were detected among those contributing more than 0.1% to the "energy" of $\mathcal{D}$. For MEEI, we detected 44 PCs in the acoustic frequency and 29 PCs in the modulation frequency subspace. This resulted in a reduced space of $44 \times 29 = 1276$ features. For PdA, the corresponding reduced space had dimensions of $53 \times 36 = 1908$. Next, the features which were more correlated to the voice pathology detection task were selected for each database, using the Maximal Relevance criterion (MaxRel). For details about the application of the MaxRel criterion on this task please refer to Chapter 3.

To extract MFCC features, each utterance was first run through the standard mel-cepstrum filterbank (using 12 filters) at a 25-ms frame interval. The cepstrum was computed and channel compensation techniques were applied according to [48]. In order to combine MFCC with mRMS features, the mean and variance of the 12 MFCC features over 10 frames were extracted, every 2 frames (a 50 ms shift). Delta features were not included since the improvement over MFCC features alone was not found to be statistically significant in [48].

The features were then fed as input to a support vector machine (SVM) classifier with a radial basis function kernel [60]. Detection-error tradeoff (DET) curves and the equal error rate (EER) were used to compare the performance of different systems on MEEI and PdA.

DET curve results for standard mel-cepstrum, mRMS and the concatenated feature vector (including both MFCC and mRMS features) are plotted in Figure 5.9 for MEEI and in Figure 5.10 for PdA. The top m mRMS features were selected for each database using 4-fold cross validation. The optimum detector based on mRMS features alone was obtained by considering the $m = 125$ most relevant features for both MEEI and PdA. As shown, the equal error rate (EER) - the point where the false alarm probability equals the miss probability - of mel-cepstrum alone is 8.47% on MEEI and 22.86% on PdA, with mRMS features yielding 6.29% on MEEI and 17.67% on PdA, and the concatenated vector resulting in 3.63% on MEEI and 12.15% on PdA (Table 5.7).

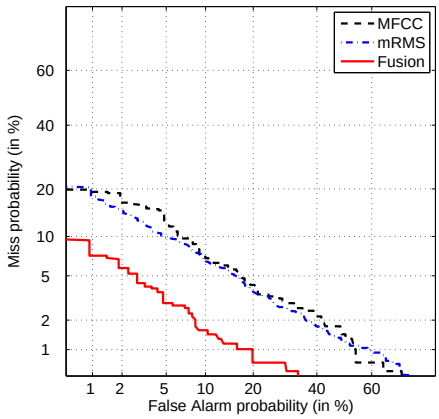


Figure 5.9: Performance of MFCC and mRMS features in MEEI.

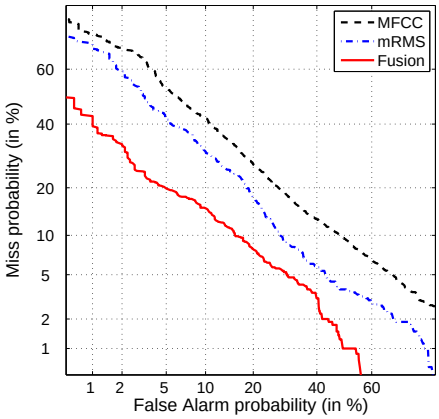


Figure 5.10: Performance of MFCC and mRMS features in PdA.

In the cross-database experiments, when training is performed on the $m = 125$ most relevant features of PdA and testing on the same mRMS features for MEEI, the EER for MFCC is 28.24%, for mRMS is 24.40% and for the concatenated features 16.87% (see Figure 5.11 and Table 5.7). When training is performed on the $m = 125$ most relevant features of MEEI and testing on the same number of mRMS features for PdA, the performance of the system significantly deteriorates. We had to consider the top $m = 450$ most relevant features - relevance estimated on MEEI - in order to capture dysphonia in PdA. In that case, the EER of mRMS is 26.07%, of MFCCs is 30.97% and of concatenated features 21.86%. Table 1 summarizes the classification scores for the different conducted experiments. The last two rows of the Table provide information for the cross-database experiment where PdA-MEEI means training on PdA and testing on MEEI and vice versa for MEEI-PdA. In brackets we note the number of the mRMS features used in each

experiment.

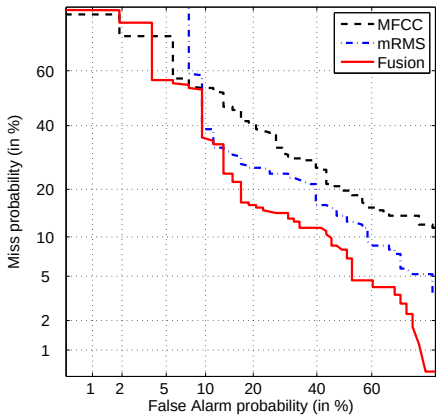


Figure 5.11: Performance of mRMS features, MFCC and their fusion when training is performed in PdA and testing in MEEI.

Table 5.7: Equal Error Rate (EER) in % for mRMS features, MFCC and both of them in MEEI and PdA.

	MFCC	mRMS	Fusion
MEEI	8.47	6.29 (125)	3.63
PdA	22.86	17.67 (125)	12.15
PdA-MEEI	28.24	24.40 (125)	16.87
MEEI-PdA	30.97	26.07 (450)	21.86

5.4.4 Discussion

Dysphonia recognition experiments on MEEI and PdA confirmed that modulation spectral features provide complementary information to MFCC. The low bands of the MFCC reflect alterations related with the mucosal waveform due to an increase of mass whereas the noisy components induced by lack of closure are modeled by the higher bands [48]. Modulation spectra on the other hand capture the amplitude envelope fluctuations evident on sustained vowel phonations [72].

Regarding cross-database experiments, features selected from PdA alone were more successful in capturing class specific information in MEEI than vice versa. A potential reason for this is that some of the normal speakers in MEEI database were recorded at different sites and over possibly different channels than the pathological subjects [72]. This makes the MEEI an easy database for classification tests. This is not the case with PdA, where the same recording conditions were

used for normal and dysphonic speakers. It follows then, that it is better to train the classifier on PdA than on MEEI.

We have simply concatenated the mean and variance of MFCC over the same segments that mRMS were estimated from; the concatenated feature vector was given as input to the SVM classifier. A better strategy, would be to combine different classifier schemes for every feature set. We ran additional experiments with MFCC and GMM classifier, as well as mRMS and GMM classifier on the same datasets for normal/pathological distinction. Configuration of MFCC with GMM classifier (the system described in [48]) was better than using MFCC with SVM - still, in all experiments MFCC plus GMM produced inferior results to the fusion of features combined with SVM. On the other hand, mRMS plus SVM configuration clearly superseded mRMS plus GMM. The reason is the large number of mRMS features and the corresponding quadratic increase of the number of parameters of GMM classifier. In a relevant work, we also explored the fusion of classifiers at the decision level instead of the fusion at the feature level [4].

5.5 Experiments on Voice Quality Assessment

5.5.1 Database

We used PdA database since the phonations it includes, have been classified according to the G parameter (grade) of the Hirano's GRASB scale. A four-point scoring system is used to rate each subject along the G dimension: 0 denotes no hoarseness (normal voice), 1 means slight dysphonia, 2 refers to moderate hoarseness, whereas 3 describes a severely hoarse voice. For the following experiments, we selected 199 subjects with normal voice (87 men and 112 women, aged 17 to 54) and 200 dysphonic subjects (74 men and 126 women, aged 11 to 76) affected by nodules, polyps, oedema, etc. Specifically, we used 124 dysphonic voices with grade $G = 1$, 71 with $G = 2$ and 5 voices with $G = 3$. Due to the very small number of subjects with a hoarseness rating equal to 3, these were joined with the subjects with a rating 2 hoarseness.

5.5.2 Methods

Projection of modulation spectral features on the principal axes with the higher energy in each subspace results in a compact set of features with minimum redundancy. We further selected features which were more relevant to the given classification task using mutual information (MI). Mutual information to the class variable was estimated twice: for the normal/dysphonic classification of phonations of sustained vowel /AH/, and for the discrimination of pathological subjects

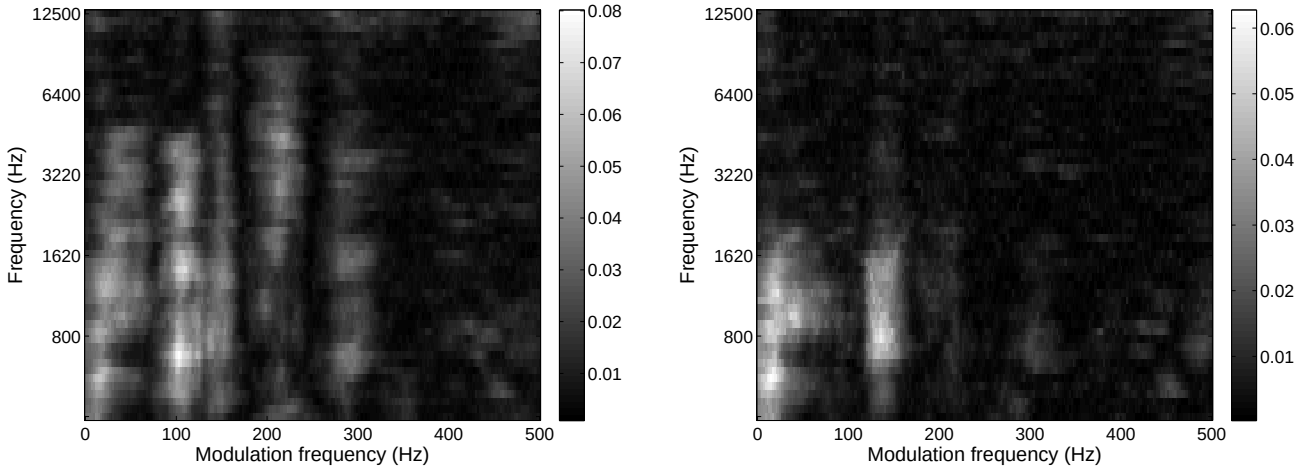


Figure 5.12: Mutual information of the original normalized modulation spectral features (a) for the normal/dysphonic classification of phonations of sustained vowel /AH/ in PdA; (b) for the classification of hoarseness in 2 grades ($G = 1$ and $G = 2$) for the dysphonic only phonations.

into two hoarseness scores. Two different feature sets were thus defined according to the sorted MI values. Fig. 5.12a depicts the mutual information of the original normalized modulation spectral features for the classification of the sustained vowel phonations in PdA in normal and dysphonic classes. Fig. 5.12b depicts the mutual information of the original normalized modulation spectral features for the classification of hoarseness of the dysphonic only phonations in 2 scores (G_1 and G_2). Modulations localized lower than $\sim$ 1600 Hz on the acoustic frequency axis seem to be more relevant; this is consistent with previous experimental results on pathological voice assessment where frequencies lower than 3000Hz led to an homogeneous discrimination between voices compared with higher frequencies [103].

We conducted some preliminary experiments on grade classification using two naive bayes (NB) classifiers built on top of these features. We had also tested SVM with radial basis function kernel in the first classification subtask in the same corpus [86]. SVM gave almost equivalent results to NB classifier but the running time was much longer. We therefore used NB classifiers in both experiments. We used leave-one-out cross-validation to select the top m features in every set; further the two classifiers were combined in a binary tree as we describe next.

5.5.3 Results

Modulation spectra were computed in a frame-by-frame basis using long windows in time (262 ms) which were shifted by 64 ms. We used Mel scale filtering with 53 bands while the size of the Fourier transform for the time-domain transformation was set to 257 (up to π). Therefore, each modulation spectrum consisted of $I_1 = 53$ acoustic frequencies and $I_2 = 257$ modulation frequencies, resulting therefore in an 53×257 "image" per frame. The modulation spectra computed in each frame were mean subtracted and then they were stacked to produce a third order tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$, where I_3 is the number of frames in the training dataset. After applying the High Order SVD algorithm, we kept the principal axes (PCs) of features contributing more than 0.1% to the "energy" of $\mathcal{D}$; i.e., the first 43 PCs in the acoustic frequency and the first 29 PCs in the modulation frequency subspace. This resulted in a reduced space of $43 \times 29 = 1247$ features. Next, the features which were more correlated to the hoarseness assessment were selected using the Maximal Relevance criterion (MaxRel). For details about the application of the MaxRel criterion on this task, please refer to Chapter 3.

Two different feature sets were defined according to the sorted MI values. The first set included the most relevant features when MI estimation also involved voices from 199 normal subjects with zero hoarseness. The second feature set was selected using the dataset of dysphonic only voices (200 subjects). We used leave-one-out cross validation to select the top m features for every NB classifier built on top of each feature set. Training and testing was based on 262ms segments; utterance classification then, relied on the median of the decisions over its segments. NB classifier built on top of the $m = 150$ most relevant features of the first set was optimum for discriminating normal (class $G = 0$) from dysphonic subjects ($G \geq 1$). Table I presents the confusion matrix from the automatic classification of the subjects into normal and dysphonic voices. This classification is compared with the original perceptual judgement in the PdA corpus. For classes $G = 1$ and $G = 2$, the optimum NB classifier was obtained by considering the top $m = 420$ most relevant features of the second set. In Table II the confusion matrix from the automatic classification of the dysphonic only subjects into two scores of grade is presented. We used a simple multi-class classification scheme by constructing a bottom-up binary tree for classification. First, an utterance is classified into normal or dysphonic class. Then, dysphonic samples were further classified into $G = 1$ and $G = 2$ classes. Table III presents the overall classification results obtained by combining the NB classifiers based on these different feature sets; a global classification rate of 73.93% was achieved. We can observe that the worse performance corresponds to the $G = 1$ class where the hoarseness of dysphonic voices has been underestimated.

Table 5.8: Confusion matrix for the automatic classification of phonations into normal ($G = 0$) and dysphonic classes ($G = 1$ and $G = 2$); average accuracy is 82.21%.

Perceptual judgement	Total Number	Classification results	
		Normal phonations	Dysphonic phonations
Normal phonations	199	**167**	32
Dysphonic phonations	200	39	**161**

Table 5.9: Confusion matrix between scores of hoarseness given by the automatic classification system of dysphonic only phonations (S-G_1, S-G_2) and their respective perceptual judgement (P-G_1, P-G_2); average accuracy is 80.5%.

Perceptual judgement	Total Number	Classification results	
		S-G_1	S-G_2
P-G_1	124	**106**	18
P-G_2	76	21	**55**

5.5.4 Discussion

In this work we have proposed a method for objective assessment of hoarse voice quality, based on modulation spectra. Using mutual information we could locate the most relevant frequency bands at the "formant zone", i.e. lower than 3000 Hz. Based on different feature sets, two NB classifiers were tested and found to be optimal in the discrimination of different classes. By combining them into a simple binary tree classification scheme, a global classification rate of 73.93% was achieved.

Future work could address additional GRASB parameters using a database of reading text. We could explore the discriminative ability of consonant classes as well in the objective assessment of different voice qualities. In addition, benchmarking against more standard approaches like those used for the automatic speaker recognition [103] could be performed.

Table 5.10: Overall confusion matrix between scores of hoarseness given by the automatic classification system (S-G_0, S-G_1, S-G_2) and the perceptual judgement of phonations (P-G_0, P-G_1, P-G_2); average accuracy is 73.93%.

Perceptual judgement	Total Number	Classification results		
		S-G_0	S-G_1	S-G_2
P-G_0	199	**167**	32	0
P-G_1	124	33	**73**	18
P-G_2	76	6	15	**55**

5.6 Discussion and Conclusions

It was shown in [86] that Modulation Spectra can be appropriately normalized in order to successfully address the detection of dysphonic voices in new, unseen, databases. However, Normalized Modulation Spectra have not been applied yet to the task of disorders classification for new databases. Currently we are looking for a new database with enough examples from each disorder in order to conduct experiments with Normalized Modulation Spectra. A very important problem in voice disorders is the quantification of the degree of voice pathology (i.e., degree of breathiness, roughness and hoarseness). The results presented in [88] using modulation spectra for quantifying hoarseness were very encouraging. As a future plan, we would like to quantify the degree of voice pathology for the other cases too, but using more databases that the one used in [88].

Moreover, regarding future plans, analysis of continuous speech samples could be used instead of sustained vowels. Acoustic features derived from continuous speech provide information about the voice source, vocal tract and articulators, shedding light on more aspects of a pathological voice quality. In that case, we expect that higher (acoustic) frequency bands in the modulation spectra would also contain highly discriminating patterns for vocal pathologies assessment. Different Time-Frequency (TF) distributions could also be used in the first stage of modulation frequency analysis instead of the STFT spectrogram, offering better resolution [67]. Also, alternative time-frequency transformations, such as decomposition based approaches, proposed in a previous study [133], could also be used.

Chapter 6

Classification of Systolic Heart Murmurs based on Reassigned Spectra

Synopsis of the Chapter

Heart sounds were recorded in children with a systolic cardiac murmur using an electronic stethoscope. A system that detected the beginning of every heart cycle, and subsequently characterized systolic murmurs, was developed. R-waves in the synchronous ECG signal were used as reference for the detection of the first heart sound. Sound signal analysis was carried out using reassigned spectrogram. Using an information theoretic criterion based on the Mutual Information between systolic murmur classes (i.e., innocent/abnormal) and features, the relevance of certain frequencies at specific time instants in the heart cycle segment could be quantified. The initial time-frequency representation was transformed to a lower-dimensional domain using higher order singular value decomposition (HOSVD). The classification system was evaluated through cross-validation experiments on a database of recordings from functional and abnormal heart murmurs of the Pediatric Cardiology Unit of the University Hospital of Heraklion. Using support vector machines (SVM) for classification, the suggested approach achieved an Equal Error Rate (EER) of $7.07 \pm 3.65\%$ and an Area Under the Curve (AUC) score of 0.9733 ± 0.0213. This performance is comparable to the reported accuracy achieved by experienced pediatric cardiologists while significantly better than that of general pediatricians.

This chapter is based upon the following publication:

- Markaki M., Germanakis I. and Stylianou Y., *Automatic classification of Systolic Heart Murmurs Based on Reassigned Spectrogram*, to be submitted to IEEE Transactions on Biomedical Engineering.

I would like to thank Dr Ioannis Germanakis (Pediatric Cardiology Unit of the University Hospital of Heraklion) for the PCG data and for his collaboration.

6.1 Introduction

Accurate and early diagnosis of the abnormalities in the cardiovascular system of children is of great importance. Almost 1% of the children suffer from an abnormality of the cardiovascular system (congenital heart disease or inherited cardiomyopathy), which results to severe morbidity and mortality if there is no early diagnosis [38]. Classic heart auscultation using a conventional stethoscope to detect abnormal heart sounds is the most common and widely recommended method to screen for structural abnormalities of the cardiovascular system [57, 92]. Detecting relevant symptoms and forming a diagnosis based on the sounds heard through a stethoscope, however, is a skill that can take years to acquire and refine [44]. It has been proven that general practitioners and pediatricians cannot accurately distinguish pathological (related to underlying heart disease) heart sounds from the innocent functional murmurs that appear to a significant percentage of children. Their diagnostic accuracy is internationally reported as low to moderate. Suboptimal clinical skills of physicians and pediatricians result in inappropriate referrals and misuse of expensive diagnostic methods such as the echocardiogram, which is the predominant noninvasive diagnostic method in cardiology nowadays [110, 73, 44].

Modern technological advances enable both high quality digital recording and reproduction of heart sounds using electronic stethoscopes as well as the remote, on-line or off-line, auscultation in children. Depending on the digital stethoscope, a sensitivity of 87-100% and a specificity of 82-98% in the discrimination of innocent and abnormal murmurs has been reported when *experienced pediatric cardiologists* evaluate digital recordings [39, 23]. These studies were based on selected recorded samples which might not be representative. In [44], the same method was applied to random samples of patients and the sensitivity of the off-line remote auscultation was evaluated as a function of the severity of heart disease in children. In this study, 96% of moderate to severe congenital heart disease and more than 92% of children with functional murmurs were accurately detected in blind, off-line auscultation by two experienced pediatric cardiologists from Greece and Germany. Since now, to our knowledge, there are no reports for systems designed and clinically tested as screening tools to detect heart diseases in school-aged children. In this age the distinction between abnormal sounds from the functional nature is of great clinical significance as most children manifest additional heart sounds (murmurs) of non significance (innocent murmurs) which should be discriminated with the more rare abnormal heart sounds (abnormal murmurs) associated with underlying cardiovascular malformations.

Normal heart sounds (HS) mainly consist of two regularly repeated consequent sounds, known

as S_1 and S_2, for every heart cycle. S_1 corresponds to the closing of the atrioventricular (inlet) valves immediately preceding the systole, while S_2 corresponds to the closing of the arterial (outlet) valves at the end of the systole. The time interval from S_1 to S_2 defines systole (heart contraction) associated with active pumping of the blood within the arterial vessels, while the S_2-S_1 space defines the diastole (heart relaxation) associated with passive refilling of heart chambers. The normal blood flow within the heart and vessels is mainly laminar and therefore silent; but if passing through abnormal communications or narrowed valves the blood flow becomes turbulent, and by causing vibration to surrounding tissues gives rise to audible additional sounds (murmurs).

Murmurs are characterized by their intensity, spectral content, temporal duration and their position in the cardiac cycle phases (systolic, diastolic, continous) [130, 44, 43]. These signals are non-stationary, since their characteristics are not stable over time [25, 71, 66]. Moreover, due to the variation in the physiology of the heart and other body parts (e.g. chest) among humans, murmurs in the same category (pathological or functional) differ among different individuals. Also, heart sounds are usually recorded under additive and convolutional noise conditions. Therefore, extracting robust features from heart sound signals is not straightforward.

Automatic diagnosis in pediatrics using computers has only recently been studied [111] with promising results. There have been a few attempts towards *automatic* discrimination between functional and abnormal murmurs. Automatic here refers to the process of features extraction which has to be relied only on a software system. Most early approaches in automatic diagnosis based on the heart sound characteristics rely on wavelet analysis for feature extraction [52, 127]. More recently, different time-frequency representations - such as Exponential Time Distribution (ETD) and Hyperbolic Time Distribution (HTD) - have been tested [109]. These non-parametric approaches are based on the magnitude spectrum of signal. Also short-term mel frequency cepstral coefficients (MFFC) have been used in representing the magnitude spectrum in a compact way [109]. The correlation of the various suggested features and representations with murmur pathology was evaluated using techniques like linear multiple regression analysis [33], or validation experiments [109]. Also neural networks and k-nearest neighbor classifiers have been suggested [26, 16, 52, 127, 109].

In this study, we develop a system for automatically discriminating functional from pathologic systolic murmurs based on the reassigned spectrogram [40] for the analysis of phonocardiographic signals, and Support Vector Machines as classifier. The short -time Fourier transform (STFT) is used as the input time-frequency representation for the reassignment algorithm; still, other time-frequency or time-scale representations could also be used [67, 9].

In Section 6.2, a general description of the database is provided along with its subsets used in the classification experiments. For this purpose, Phonocardiogram (PCG) recordings from the Pediatric Cardiology Unit, Dpt of Pediatrics, University Hospital of Heraklion, Crete were considered. In Section 6.3, the time - frequency analysis framework is briefly described and examples of PCG signals recorded by normal and pathological subjects are presented. In the same section, the relevance of the high resolution time-frequency representation of PCG signals to the discrimination of systolic heart murmurs to innocent and pathological, is evaluated. Relevance can be objectively defined through the Mutual Information (MI) measurement based on Information Theory [22] (cf Chapter 3, on the Mutual Information (MI) estimation procedure). The large dimensionality of the time-frequency representation poses a serious problem to the classification. We approach the dimensionality reduction of the time - frequency representations in the framework of multilinear algebra using higher order singular value decomposition (HOSVD) [68] (cf Chapter 3, on the redundancy reduction). We present the pattern classification algorithm and the performance analysis measures used in the paper in Section 6.4.1. Finally, conclusions are drawn and future directions are indicated in Section 6.5.

6.2 Data Collection

6.2.1 Subject Population

Pediatric cardiology outpatients referred either for murmur evaluation or followed-up for known heart disease in the Outpatient Pediatric Cardiology Clinic of the University Hospital of Heraklion, Crete, which had undergone both a complete two-dimensional echocardiographic study and a digital phonocardiogram recording during a 18-month period (1.1.2008-30.6.2009), were eligible for enrollment in the study. Anonymized data storing and off-line analysis of the PCGs for teaching and research purposes has been approved by the Hospital Ethics Committee and consent was obtained from patients' parents.

6.2.2 Data Acquisition - Equipment

Following a detailed echocardiographic study (using Vivid 3 Expert, GE ultrasound system and age appropriate transducers of 2.5-5MHz) which was performed and interpreted by a single academic pediatric cardiologist (second co-author), a digital phonocardiogram was also recorded. For the latter, a sensor based electronic stethoscope with incorporated 3 lead ECG was used (TheStethoscope, Welch Allyn-Meditron, AS, Norway). Patients were lying supine and four

recordings of 6 sec duration each were performed, corresponding to the apical, lower (4th intercostal space) left sternal border (LLSB), upper (2nd intercostal space) left (ULSB) and upper right sternal border (URSB) as well as a jugular fossae recording (JGLR). Digital acoustic data (with a sampling rate of 44,100 KHz, 16 bit dynamic resolution) and ECG signals, allowing for a precise temporal classification of cardiac sounds, were transferred and stored as wav files, in a personal laptop computer using the designated software (Meditron Analyzer 4). The digital phonocardiograms were assigned to corresponding patients by a random ID number. A pediatric cardiologist ((2nd co-author), by using a set of good quality headphones (MX 450, Sennheiser GmbH&Co.KG with a frequency response of 18Hz-21kHz), blindly validated the available digital phonocardiograms at a later time, regarding: a) the presence of a murmur, b) murmur timing (systolic, diastolic, continuous), quality (innocent, regurgitant, ejection), intensity (grades 1 to 3 in increasing order), site of maximum intensity (p.m.) and radiation sites, and c) the presence of additional abnormal auscultatory findings as systolic clicks and second heart sound abnormalities (fixed wide split).

If almost all digital phonocardiograms of a given case were of too low quality to be accurate interpreted by the pediatric cardiologist, this case was excluded from the original phonocardiogram database. Cases having both detailed final echocardiographic diagnosis and precise off-line description of auscultatory findings were only selected.

6.2.3 Phonocardiograms Selection

The first 25 consecutive patients of the original database, with innocent murmurs and an otherwise normal echocardiographic study were included in the analysis as control group (innocent murmur group). Similarly, the first 25 consecutive patients of the original database, with abnormal murmurs (systolic or continuous with extension in the systolic period) and a definitive echocardiographic diagnosis of various forms of congenital heart disease, were included in the analysis as cases group (abnormal murmur group). The study database therefore consisted of 50 cases with systolic murmurs, with balanced representation of innocent and abnormal systolic murmurs (Tables 6.1 and 6.2). As a total of 5 recording positions were available for evaluation for each patient, a total of 250 digital phonocardiograms were eligible for further analysis.

In the present study, only the first three consecutive phonocardiograms for each case were used, whenever possible, corresponding to the apical, lower and upper left parasternal recordings. In case of suboptimal recording conditions in a given position (sound artifacts as validated during off-line reproduction), the recording position was excluded and the subsequent recording position

Table 6.1: Abnormal murmur database

Case No	Age (yr)	Diagnosis	Severity	murmur	Apex	LLSB	LUSB	RUSB	JGLR
1	6.5	AVSD / PS	mild	ejection	1	1			1
2	9.5	AOS	mild	ejection	1	1	1		
3	1.0	PDA	mild	continuous	1	1	1		
4	3.9	ASD	moderate	ejection	1	1	1		
5	10.8	ASD	moderate	ejection	1	1	1		
6	0.3	VSD/ASD	moderate	regurgitant	1	1	1		
7	10.8	L-TGA/PS/TI	severe	ejection,regurgitant	1	1	1		
8	9.3	VSD	moderate	regurgitant	1	1	1		
9	15.7	AOS	mild	ejection	1	1	1		
10	4.8	MVP/MI	moderate	regurgitant	1	1	1		
11	2.3	ASD	moderate	ejection	1	1	1		
12	0.4	PDA	mild	continuous	1		1	1	
13	10.0	PS	mild	ejection	1	1	1		
14	6.4	AS suprav.	mild	ejection	1	1	1		
15	5.8	HCM	mild	ejection	1	1	1		
16	9.7	AOS	mild	ejection	1	1	1		
17	3.3	AOS	mild	ejection	1	1	1		
18	0.1	VSD	mild	regurgitant			1	1	1
19	11.9	HCM	mild	ejection	1	1	1		
20	4.6	AOS	mild	ejection	1	1	1		
21	8.5	PS	mild	ejection	1	1	1		
22	6.1	AOS/COA	mild	ejection		1	1	1	
23	0.1	TOF	severe	ejection	1	1	1		
24	3.6	post TOF	mild	ejection	1	1	1		
25	5.2	VSD	mild	regurgitant			1	1	1

Abbreviations: AVSD: atrioventricular septal defect; AOS: aortic stenosis (all types); PS: pulmonary valve stenosis; VSD: ventricular septal defect; ASD: atrial septal defect; HCM: hypertrophic cardiomyopathy; COA: coarctation of aorta; TOF: tetralogy of Fallot; TGA: transposition of great arteries; PDA: persistent arterial duct; MVP/MI: mitral valve prolapse - insufficiency; TI: tricuspid valve insufficiency.
Recording positions: Apex, LLSB (left lower sternal border), **LUSB, RUSB** (left and right upper sternal border), and **JGLR** (jugular fossae).

was used in the analysis. In this way, the availability of 3 auscultation sites recording the same event was assured. As a wide variety of congenital heart defects was included, having variable sites of abnormal murmur maximum intensity points and variable intensity, the selection process of the three auscultation sites, would also result in the detection of low intensity murmurs not radiating to the whole chest area. Tables 6.1 and 6.2 present the auscultatory recording positions which were used in the current work.

Table 6.2: Innocent murmur database

Case No	Age (yr)	murmur intensity*	Apex	LLSB	LUSB	RUSB	JGLR
1	10.1	3	1	1	1		
2	7.0	1	1	1	1		
3	12.2	2		1	1		1
4	4.5	3	1	1		1	
5	7.6	2	1	1	1		
6	3.8	3	1		1	1	
7	11.8	1	1	1	1		
8	3.9	2	1	1	1		
9	4.3	3	1	1		1	
10	3.0	2	1	1	1		
11	3.6	2		1	1	1	
12	3.1	3	1	1	1		
13	4.5	2		1	1	1	
14	5.2	2			1	1	1
15	6.4	3	1	1	1		
16	0.3	1			1	1	1
17	5.3	1	1	1	1		
18	8.4	2	1	1	1		
19	8.5	1	1	1	1		
20	8.1	2	1	1	1		
21	8.5	2	1	1	1		
22	8.4	2	1	1	1		
23	8.6	3	1	1	1		
24	9.0	2	1	1	1		
25	8.0	2	1	1	1		

*murmur maximum intensity in a 3 degree scale based on off-line assessment of digital phonocardiograms by pediatric cardiologist

Recording positions: Apex, LLSB (left lower sternal border), **LUSB, RUSB** (left and right upper sternal border), and **JGLR** (jugular fossae).

6.3 Methods

6.3.1 Automatic Preprocessing of PCG recordings

Software was developed for the automatic detection and analysis of heart cycles and murmurs. The PCG-synchronous ECG signals were used as reference for detecting heart sounds since the R-peaks in ECG coincide with the beginning and end points of heart cycles. Both ECG and PCG signals were down sampled to the same sampling frequency. The ECG signal was band-pass filtered with cut-off frequencies of 10 and 25 Hz and the R-waves were detected using an envelope based detection algorithm [32]. Shannon energy was used for the envelope estimation

of the ECG signal $y[n]$, defined as:

$$E_S[n] = -y[n]^2 \log(y[n]^2). \tag{6.1}$$

The Shannon energy, $E_S(n)$, emphasizes the medium intensity sounds, attenuates the effect of low intensity components much more than high intensity components and eliminates noise [1]. An algorithm based on Shannon energy and local maximum peaks within a sliding window was used to locate R-peaks of the QRS-complex. The R-R intervals were used then as a reference for heart cycle detection and segmentation of the PCG signals [32, 31] (see Figure 6.1 below).

After segmentation of the ECG-synchronous PCG signal, the individual heart cycles have different lengths depending on the cardiac rhythm. Heart rates in children are quite variable depending on their age and physiological state (fever, fear, activity level etc). The length of the systole however is quite stable as increased heart rate is achieved physiologically at the cost of reduced diastolic (relaxing) interval.

The purpose of the developed objective diagnostic method, is to detect abnormal characteristics in systolic murmurs only, such as changes in their frequency content, timing and intensity. This is an important diagnostic issue, as both innocent and some of the abnormal murmurs are systolic: all diastolic or continuous murmurs are by definition abnormal requiring further echocardiographic evaluation, therefore they were not included in the analysis [19]. Additional sounds (clicks) and pathological variations in the intensity or timing of heart sounds (S1 and S2) also characterize most of the anatomical abnormalities in the cardiovascular system of children (congenital heart diseases) [43]. Systolic murmurs might extend up to and merge with the heart sounds, which at times are difficult to be detected in the presence of loud or holosystolic murmurs. By including the early phase of the systole, some pathological conditions with diagnostic findings early in the systole would not be missed [32, 33, 31]. In arterial (outlet) valvular stenosis, for example, early systolic clicks (brief duration sounds close to S1) are important auscultation findings which should not be ruled out of the analysis [32, 31]. In [33, 31] the distinguishing variables for the detection of cardiac pathology included the interval from the end of the first heart sound (S1) and the beginning of the systolic murmur, and the standard deviation of the interval from the end of S1 to the maximum intensity of the murmur. S2 on the other hand, consists of two components, aortic (A2) and pulmonary (P2) valve closure sounds; normally A2 is of higher intensity than P2, while P2 shows a characteristic delay only during inspiration. The interval between the two components of S2 and its variation with respiration has also important

clinical significance [43, 33, 31].

Consequently, the time-frequency analysis should take jointly into account the first heart sound (S_1), the whole systolic phase and the second heart sound (S_2) - in at least three consequent heart cycles for detection of any respiratory variation [43]. According to [32, 31], if Y denotes the length of the systole (in sec) and X the heart rate (in beats per minute), then:

$$Y = 0.351 - 0.001 * X. \tag{6.2}$$

For example, if the heart rate, X, varies between 60 and 160 beats per minute, then the systole length, Y, will vary between 291 and 191 ms respectively. We have chosen then to analyze a fixed-length segment comprising of the first 400 ms from the beginning of the heart cycle. For every recording, five consequent heart cycles were selected. Figure 6.1 depicts the phonocardiographic signal (solid) and the envelope of electrocardiographic signal (dash) of an 10-years old with innocent early to midsystolic murmur. Five consequent heart cycles are shown, with the maxima of ECG energy, i.e., R-peaks, denoted by stars pointing to the beginning of each heart cycle. A 400ms segment beginning at the first heart cycle is presented; it includes S_1, the systolic murmur (SM) and S_2. It was found that this 400ms segment always included these heart sounds; seldom, in bradycardia, a short part of S_2 might be missed.

PCG signals were band-pass filtered with cut-off frequencies of 40 and 1100 Hz as recommended by [32, 31]. The 3rd order Butterworth type high-pass and low-pass filters were used [99]. The amplitude of each signal was scaled by the absolute maximum of each PCG recording according to [99]. Thus the resulted waveform showed the relative intensities of the PCG as the ear recognizes them (see Figure 6.1).

6.3.2 Time-Frequency representation of heart sounds

To follow the time varying characteristics of the heart sounds (i.e., S_1, SM, S_2) a time-frequency representation of these sounds is required. The most common time-frequency representation is that obtained by the short time Fourier Transform (STFT), which is referred to as spectrogram. A spectrogram contains just a few cross-terms arising from interference phenomena; however this is obtained at the expense of poor concentration and resolution of the individual components of multi-component signals [41]. As it is usually stated, a spectrogram is constrained by the trade-off between resolution in time and frequency. The reassignment method which is briefly described in Chapter 2, was specifically designed to overcome this trade-off (see [9, 40, 41] for

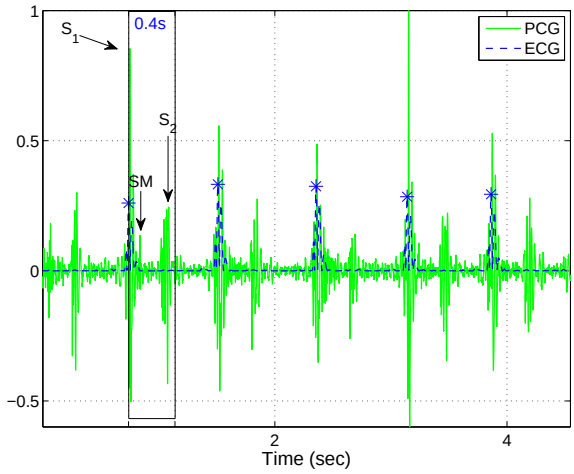


Figure 6.1: Phonocardiographic signal (solid) and envelope of electrocardiographic signal (dash) of an 10-years old with innocent early to midsystolic murmur. The auscultation area is the left lower sternal border (LLSB) where the murmur is best heard, although it radiates to all the positions. Five consequent heart cycles are shown with the maxima of ECG energy (stars coincide with R-peaks) pointing to the beginning of each heart cycle. A 400ms segment beginning at the first heart cycle is shown, including S_1, the systolic murmur (SM) and S_2.

more details on the reassignment method).

6.3.3 Spectral Patterns and Time-Intervals in Innocent and Pathological Murmurs

The selection of a low sampling frequency significantly reduces the computational complexity of reassignment spectrogram estimation. Since a cut-off frequency of 1100 Hz is recommended for low-pass filtering the PCG [32], PCG (as well as synchronous ECG) signals were down sampled to a sampling frequency of 2205 Hz.

Figure 6.2 depicts a case of innocent systolic murmur of a 10-years old (case 1 in Table 6.2), with murmur intensity equal to 3 in a 3-degree scale (as evaluated by the pediatric cardiologist I.G.). Figures 6.3a and 6.3b show examples of systolic murmurs in children with different heart pathologies: atrial septal defect and (mild) aortic stenosis, cases 5 and 9 in Table 6.1, respectively. The auscultation area in all cases was the left lower sternal border (LLSB). The upper panel in these figures shows two frequency contours. The blue solid contour describes the frequency corresponding to the maximum sound intensity in every time instant t:

$$F_{max}(t) = arg \max_{\nu}(S^{(r)}(t,\nu)) \tag{6.3}$$

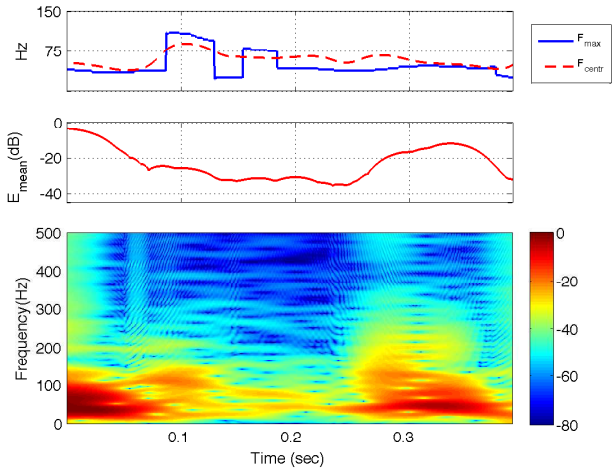


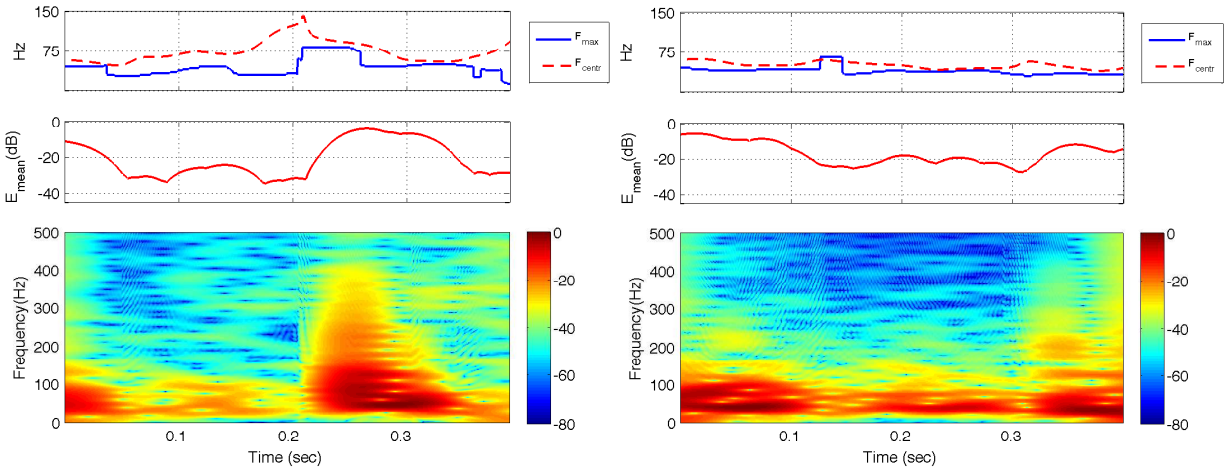
Figure 6.2: Upper panel: F_{max} (blue) and F_{centr} (red dashed) frequency contours (in Hz). Middle panel: mean energy contour (relative sound intensity in dB). Lower panel: energy (relative sound intensity in dB) of the reassigned spectrogram of the PCG (shown in Figure 1) of an 10-years old with innocent early to midsystolic murmur. The auscultation area is the left lower sternal border (LLSB) and the murmur intensity is 3.

On top of the F_{max} frequency contour, the frequency centroid contour, F_{centr}, is superposed (red dashed contour), estimated as a weighted average of the frequencies in every time instant, t:

$$F_{centr}(t) = \frac{\sum_{\nu} \nu S^{(r)}(t,\nu)}{\sum_{\nu} S^{(r)}(t,\nu)} \tag{6.4}$$

where frequencies are weighted by the corresponding sound intensities. The middle panel of each figure shows the mean spectral amplitude $S^{(r)}(t,\nu)$ over the frequencies $\nu \in 40-100$ Hz (relative sound intensity in dB). The time function describing the mean amplitude may be used to determine the locations of S_1 and S_2 according to [32, 31] and references therein. S_1 corresponds to the highest peak within the time-interval between 0.05R-R and 0.2R-R, where R-R denotes the interval between two consequent R-peaks of ECG, i.e., one heart cycle. S_2 detection can proceed in a similar way within the interval from 1.2R-T to 0.6R-R, where the peak of T-waves of the ECG are used as references too [32, 31]. The lower panel is the reassigned spectrogram amplitude $S^{(r)}(t,\nu)$. The intensities in the time-frequency distributions were scaled by the maximum intensity in the spectrogram of each PCG segment. The relative intensities in the decibel scale were calculated then, with the value of 0 dB corresponding to the maximum intensity; the corresponding intensity scale in dB is shown on the right side.

Fig. 6.4a shows the mean energy distribution (in dB) of the reassigned spectra $S^{(r)}(t,\nu)$ of 400

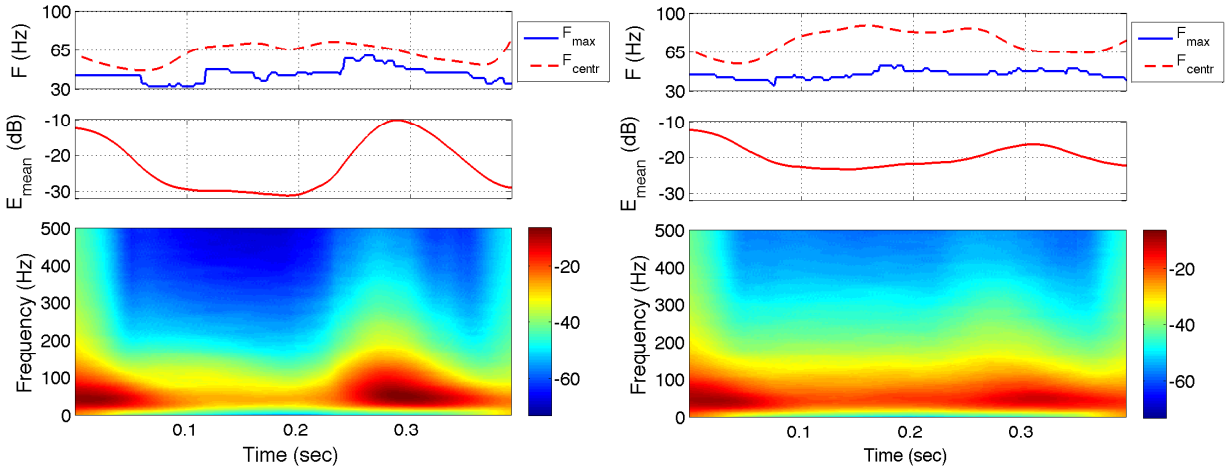


(a) (b)

Figure 6.3: Upper panel: F_{max} (blue) and F_{centr} (red dashed) frequency contours (in Hz). Middle panel: mean energy contour (relative sound intensity in dB). Lower panel: energy (relative sound intensity in dB) of the reassigned spectrogram of the PCG of (a) a 11-years old with ejection murmur and wide split P2 (murmur intensity 2). Final diagnosis : hemodynamic significant Atrial Septal Defect (inter-atrial communication with volume overload of right heart chambers and abnormally increased blood flow through an otherwise normal pulmonary valve). (b) a 16-years old with a systolic click and ejection mrm (intensity 2) suprasternal with small interatrial communication. Final diagnosis : bicuspid aortic valve, mild aortic stenosis. On auscultation, the typical systolic click was best heard at the suprasternal area. The auscultation area was the left lower sternal border in both cases (LLSB).

ms long segments from the beginning of each heart cycle in the PCG of 25 children with innocent murmurs and no heart pathology present as confirmed by echocardiogram (3 recordings with 5 consequent heart cycles per recording). All spectra have been normalized by dividing them with their maximum value prior to logarithm estimation and averaging. Thus the maximum energy value of each spectrum is 0 dB.

Spectral representations of PCG segments with innocent systolic murmurs are quite similar to each other. The musical murmur caused by harmonic movements of the heart or the vasculature, is usually visualized as a well-defined area in the spectrogram. Innocent systolic murmurs appear to have a lower peak frequency, below 200 Hz, and shorter duration than abnormal murmurs. An innocent systolic murmur always fades before the second heart sound, S_2. Thus, the energy appears quite localized in time at the two heart tones, S_1 and S_2 (middle and lower panel), at frequencies up to 200 Hz, whereas in the systolic phase (i.e., between S_1 and S_2) significantly lower energy values appear at frequencies lower than 100 Hz, at the early to mid-systolic phase.



(a) Innocent murmurs

(b) Pathological murmurs

Figure 6.4: Upper panel: F_{max} (blue) and F_{centr} (red dashed) frequency contours (in Hz). Middle panel: mean energy contour (relative sound intensity in dB). Lower panel: mean values for the energy (relative sound intensity in dB) of the reassigned spectra of the PCG from 25 subjects with (a) innocent systolic murmurs (3 recordings with 5 consequent heart cycles per recording), (b) pathological systolic murmurs (3 recordings with 5 consequent heart cycles per recording). All spectra have been normalized by dividing them with their maximum value prior to logarithm estimation and averaging.

The higher the velocity of abnormal blood flow as detected in doppler echocardiography, the more intensive the murmur and the wider the frequency scale. This phenomenon is clearly visible in the reassigned spectrogram in the case of a heart pathology. Fig. 6.4b shows the mean energy distribution (in dB) of the reassigned spectra $S^{(r)}(t,\nu)$ of 400 ms long segments from the beginning of each heart cycle in the PCG of 25 children with pathological murmurs and various heart pathologies as validated by echocardiogram (3 recordings with 5 consequent heart cycles per recording). All spectra have been normalized prior to averaging, following the same process as described above. The cluster around the second heart tone, S_2, is less intense due to the inhomogeneity of the spectra of pathological murmurs. The energy appears uniformly spread along the whole systole length, with larger values, extending at higher frequencies compared to the energy distribution of the innocent murmurs.

6.3.4 Relevance of Reassigned Spectral Features

As previously stated, systolic heart murmurs are non-stationary signals, with varying characteristics among different individuals even within the same class (innocent or abnormal). Also,

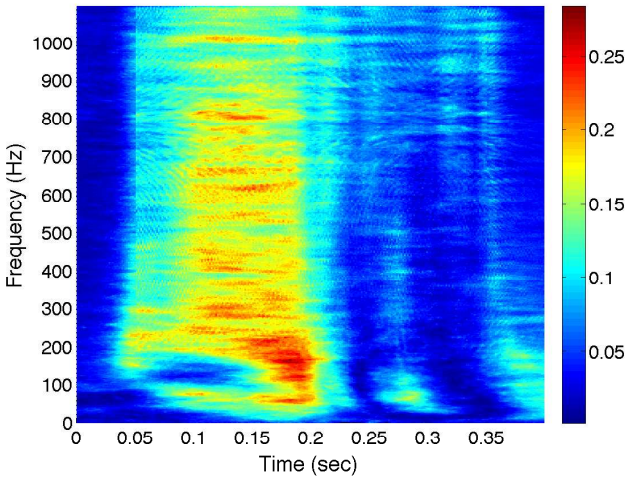


Figure 6.5: Relevance - estimated through mutual information and measured in bits - of the reassigned spectral features of the PCG for distinction of innocent from abnormal murmurs.

phonocardiographic (PCG) recordings are usually made under various additive and convolutional noise conditions. It is desirable then to extract robust against these conditions features from the PCG signals. In this section we evaluate the relevance of the features of the reassigned spectrogram of PCG for the systolic murmur pathology detection task. Relevance is defined as the mutual information (MI) $I(x_j; c)$ between feature x_j and class c (cf Chapter 3). For estimation of $I(x_i; x_j)$ we quantized the continuous space of spectral features by defining b discrete bins along each axis. An adaptive quantization (variable bin length) is adopted so that the bins are equally populated and the coordinate invariance of the MI is preserved [123]. For the data discretization involved in MI estimation, we set the number of discrete bins along each axis to $b^* = 8$ following the procedure described in [123] (cf Chapter 3).

Figure 6.5 depicts the relevance, i.e., the mutual information (measured in bits) of the reassigned spectra features of the PCG for the distinction of innocent from abnormal murmurs. The time interval in the horizontal axis covers the first 400 ms of the heart cycle. It is noteworthy that the prominent features (having MI higher than 0.15 bits) as extracted through this automatic process coincide with the empirical remarks of pediatric cardiologists [44, 45, 43]. More specifically, in [33] the authors analyzed measurements of time-intervals and spectra of the systolic murmurs using stepwise logistic regression analysis to discriminate physiological from pathological signals. They concluded that the distinguishing variables were:

- the interval from the end of the first heart sound (S_1) and the beginning of the systolic murmur; comparing Figures 6.4a, 6.4b and 6.5, it is the interval comprising the time

instants around 100 ms from the beginning of the heart cycle. Innocent murmurs never begin immediately after S_1. In Figure 6.5, the additional clue is the frequency localization of this characteristic between 50 and 100 Hz.

- respiratory variation of the splitting of the second heart sound; this is not depicted in Figure 6.5 since mutual information estimation was based on the comparison of segments out of single heart cycles, not on sequences of heart cycles.

- intensity of the systolic murmur; as seen from Figures 6.4a, 6.4b, this term corresponds to the spectral energy between 50 and 250 ms from the beginning of the heart cycle. Up to 140 ms (early to mid-systolic phase), there is a lower frequency limit of 150 of 180 Hz which progressively lowers at the next time instants (late systolic phase until the beginning of the second heart sound S_2). This comment is also related to the screening criteria presented in [99]: if a systolic murmur contained intensive frequency components of over 200 Hz and its length accounted for over 80% of the whole systolic duration, it was found to be pathological.

- standard deviation of the interval from the end of S1 to the maximum intensity of the murmur; this clue also could not be depicted in Figure 6.5 since mutual information estimation was based on the comparison of segments out of single heart cycles, not on sequences of heart cycles.

In short, the high resolution of reassigned spectral representation yields quite distinctive patterns depending on the type and the severity of heart pathology allowing thus an innocent/pathological murmur distinction. The disadvantage of this representation is its high dimensionality; moreover, the large number of most relevant features, as Figure 6.5 shows, excludes the use of a feature selection scheme. In the following section we map the time-frequency features to a lower-dimensional domain according to a technique described in Chapter 3, on redundancy reduction; for reasons of clarity, we repeat some of the related equations below.

6.3.5 Multilinear Analysis of Time-Frequency Features

Every PCG segment is represented in the time-frequency plane as a two-dimensional matrix $\in \mathbb{R}^{I_1 \times I_2}$, where I_1 and I_2 correspond to the frequency and time dimensions, respectively. Let I_3 denote the number of signal segments contained in the training set. The mean value is computed over I_3, and it is subtracted from all the spectra in the training set. The zero-mean spectra

are then stacked, creating the data tensor $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. A generalization of Singular Value Decomposition (SVD) algorithm to tensors referred to as Higher Order SVD (HOSVD) [68] enables the decomposition of tensor $\mathcal{D}$ to its mode$-n$ singular vectors:

$$\mathcal{D} = \mathcal{S} \times_1 \mathbf{U}_\nu \times_2 \mathbf{U}_t \times_3 \mathbf{U}_s \tag{6.5}$$

where $\mathcal{S}$ is the core tensor with the same dimensions as $\mathcal{D}$; $\mathcal{S} \times_n \mathbf{U}^{(n)}$, $n = 1, 2, 3,$ denotes the $n-$mode product of $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ by matrix $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$. For $n = 2$ for example, $\mathcal{S} \times_2 \mathbf{U}^{(2)}$ is an $(I_1 \times I_2 \times I_3)$ tensor given by

$$\left(\mathcal{S} \times_2 \mathbf{U}^{(2)}\right)_{i_1 i_2 i_3} \stackrel{\text{def}}{=} \sum_{i_2} s_{i_1 i_2 i_3} u_{i_2 i_2}. \tag{6.6}$$

$\mathbf{U}_\nu \in \mathbb{R}^{I_1 \times I_1}$, $\mathbf{U}_t \in \mathbb{R}^{I_2 \times I_2}$ are the unitary matrices of the corresponding subspaces of frequency and time; $\mathbf{U}_s \in \mathbb{R}^{I_3 \times I_3}$ is the samples subspace matrix. These $(I_n \times I_n)$ matrices $\mathbf{U}^{(n)}$, $n = 1, 2, 3,$ contain the n-mode singular vectors (SVs):

$$\mathbf{U}^{(n)} = \left[U_1^{(n)}\ U_2^{(n)}\ \ldots\ U_{I_n}^{(n)}\right]. \tag{6.7}$$

Each matrix $\mathbf{U}^{(n)}$ can directly be obtained as the matrix of left singular vectors of the "matrix unfolding" $\mathbf{D}_{(n)}$ of $\mathcal{D}$ along the corresponding mode [68]. Tensor $\mathcal{D}$ can be unfolded to the $I_1 \times I_2 I_3$ matrix $\mathbf{D}_{(1)}$, the $I_2 \times I_3 I_1$ matrix $\mathbf{D}_{(2)}$, or the $I_3 \times I_1 I_2$ matrix $\mathbf{D}_{(3)}$. The n-mode singular values correspond to the singular values found by the SVD of $\mathbf{D}_{(n)}$.

The contribution $\alpha_{n,j}$ of the j^{th} n-mode singular vector $U_j^{(n)}$ is defined as a function of its singular value $\lambda_{n,j}$:

$$\alpha_{n,j} = \lambda_{n,j} \,/\, \sum_{j=1}^{I_n} \lambda_{n,j} \tag{6.8}$$

By setting a threshold in the contribution of each singular vector, the R_n with $n = 1, 2$ singular vectors (SVs) can be retained for which the contribution exceeds that threshold. Thus, the truncated matrices $\hat{\mathbf{U}}^{(1)} \equiv \hat{\mathbf{U}}_\nu \in \mathbb{R}^{I_1 \times R_1}$ and $\hat{\mathbf{U}}^{(2)} \equiv \hat{\mathbf{U}}_t \in \mathbb{R}^{I_2 \times R_2}$ are obtained. Time-frequency representations $\mathbf{B} \equiv \left|S^{(r)}(\nu, t)\right| \in \mathbb{R}^{I_1 \times I_2}$ extracted from phonocardiographic signals are projected on $\hat{\mathbf{U}}_\nu$ and $\hat{\mathbf{U}}_t$ according to [68]:

$$\mathbf{Z} = \mathbf{B} \times_1 \hat{\mathbf{U}}_\nu^T \times_2 \hat{\mathbf{U}}_t^T = \hat{\mathbf{U}}_\nu^T . \mathbf{B} . \hat{\mathbf{U}}_t \tag{6.9}$$

where $\mathbf{Z}$ is an $(R_1 \times R_2)-$matrix, and R_1, R_2 denote the number of retained SVs in the frequency and time subspace, respectively.

The time-frequency representations can be approximated then in a lower-dimensional space producing a compact feature set suitable for classification. The number of retained components (or SVs) in each subspace can be determined by analyzing the "discriminative" contribution of each component. By including only the components whose contribution is larger than a threshold, we proceed to compute the cross-validation classification error (EER) as a function of this threshold in order to determine the optimal number of components.

6.4 Results

6.4.1 Pattern Classification and Performance Analysis

A binary classification task was defined that exploits the patterns of energy distribution in PCG spectra that discriminate innocent from pathological systolic heart murmurs.

Classification performance was computed when vector components were selected based on maximum contribution (maxContrib) (eq.3.4) criteria. Pattern classification was carried out using Support Vector Machine (SVM) classifiers. SVM find the optimal boundary that separates two classes maximizing the margin between separating boundary and closest samples to it (support vectors) [60]. In this work, SVMlight [60] with a Radial-Basis-Functions kernel was used.

A 5-fold stratified cross-validation was used, which was repeated 25 times. In every run, the classifier was trained on the 80% of subjects of both classes, then tested using the remaining 20%. Training and testing was based on 400ms segments from one heart cycle each; classification of PCG recordings was based on the median of the SVM decisions over the segments from five consequent heart cycles within this recording.

The system performance was evaluated using the ROC curve as well as the detection error trade-off curve (DET) between false rejection rate (or miss probability, equal to one minus sensitivity) and false acceptance rate (or false alarm probability, equal to one minus specificity) [94]. DET curves present more clearly than Receiver Operating Characteristic (ROC) curves the performance of the different assessment systems at the low error operating points [94]. The rates of each type of errors depend upon the value of a threshold. The optimal detection accuracy (DCF_{opt}) occurs when threshold is set such that the total number of errors is minimized. DCF_{opt} reflects performance at a single operating point on the detection error trade-off (DET) curve. The Equal Error Rate (EER) refers to the point at the DET curve where the false-alarm probability

equals the miss probability. We report on the average of DCF_{opt}, EER, and area under the ROC curve (AUC) along with their corresponding 95% confidence intervals as these were estimated from the 25 runs. We also report on DCF_{opt}, the respective miss and false alarm probabilities and AUC as these were estimated from the average DET and ROC curves. These rates are worse than the average rates since they correspond to a single threshold T set for every run.

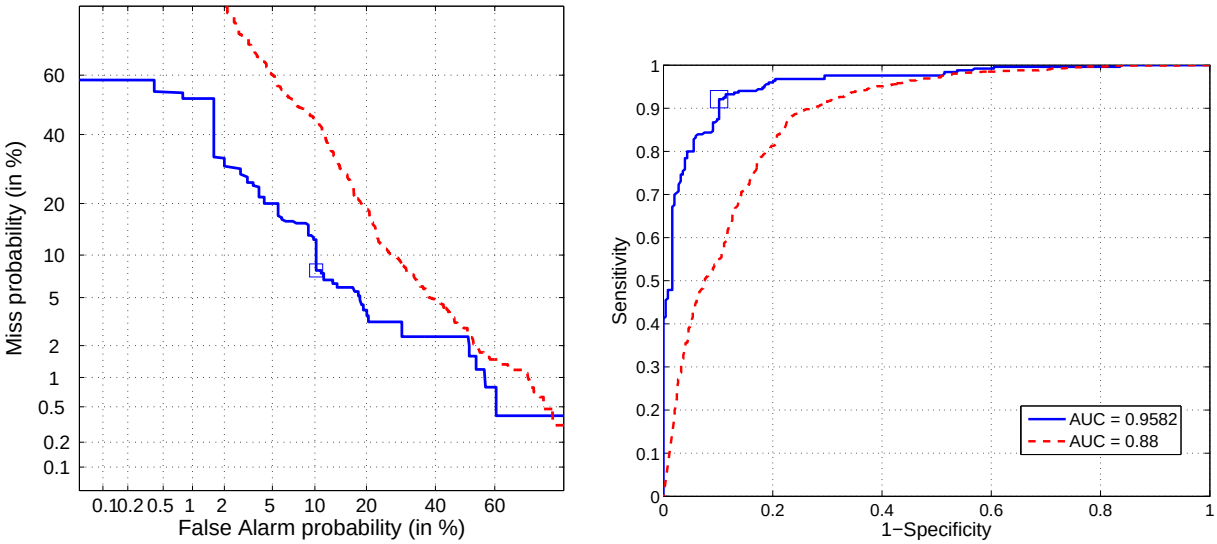
6.4.2 Feature Extraction and Classification

Reassignment spectra were computed using the Time-Frequency Toolbox [10] throughout all experiments. For the computation of the singular matrices for HOSVD, a random subset of 15 innocent and 15 pathological murmurs was selected once. Using three recordings from each subject, and considering segments of five consequent heart cycles per recording for the computation of PCG spectra, 15 PCG spectral matrices of dimension 442×294 each, were generated per subject. Stacking the $15 \times 30 = 450$ spectra matrices for all the recordings in the above subset, produced the data tensor $\mathcal{D} \in \mathbb{R}^{442\times294\times450}$. Before applying HOSVD, the mean value of the tensor was computed and then subtracted from the tensor.

The singular matrices $\mathbf{U}^{(1)} \equiv \mathbf{U}_\nu \in \mathbb{R}^{442\times442}$ and $\mathbf{U}^{(2)} \equiv \mathbf{U}_t \in \mathbb{R}^{294\times294}$ were directly obtained by SVD of the "matrix unfoldings" $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$ of $\mathcal{D}$, respectively. HOSVD is the most costly process in our system but it is performed only once. HOSVD consists of the SVD of the two data matrices $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(2)}$, with dimension $N \times k$ each. Note that the computational complexity of SVD transform is $O(Nk^2)$; N is either the frequency dimension or the time dimension of the reassigned spectrogram and k is the product of the time or the frequency dimension, respectively, multiplied by the size of the training dataset.

The singular vectors which exceeded a predetermined contribution threshold were retained in each mode (eq. 6.8), resulting in truncated singular matrices $\hat{U}_\nu$ and $\hat{U}_t$. Features were projected on these truncated orthonormal axes $\hat{U}_\nu$ and $\hat{U}_t$ according to eq. (6.9). The resulting matrices $\mathbf{Z} \in \mathbb{R}^{R_1\times R_2}$ were subsequently reshaped into vectors before SVM classification. We computed the cross-validation classification error (EER) for various contribution thresholds in order to determine the optimal dimensions R_1, R_2; the best system used $[20\times19]$ dimensions (380 features) which corresponded to a threshold of 0.02%.

Table 6.3 provides the average of classification score in terms of AUC, DCF_{opt} (with corresponding P_{miss} and P_{false}), EER and the corresponding 95% confidence intervals as these were estimated from the 25 runs. The scores per heart cycle and per recording (five consequent heart cycles selected offline) are provided. Figures 6.6a and 6.6b present the DET and ROC curves,



(a) (b)

Figure 6.6: (a) Average DET curves of 25 cross-validation runs using SVM based on one heart cycle (red dashed) or five heart cycles segments (blue solid line). Performance scores according to the average DET curve for one recording are: $DCF_{opt} = 90.4\%$, $P_{miss} = 7.89\%$ and $P_{false} = 10.18\%$ (blue square). (b) Average ROC curves of 25 cross-validation runs using SVM based on one heart cycle (red dashed) or five heart cycles segments (blue solid line). Area under the curve (AUC) scores according to the average ROC curves are: $AUC = 0.88$ for one heart cycle, and $AUC = 0.9582$ for one recording (5 consequent heart cycles). Also, the best classification score for one recording corresponds to a sensitivity of 92.11% and a specificity of 89.82% (blue square).

Table 6.3: Average performance scores with 95% confidence intervals for discrimination of innocent and pathological systolic murmurs based on one or five heart cycles (one recording)

one heart cycle			one recording				
AUC	DCF_{opt} (%)	EER (%)	AUC	DCF_{opt} (%)	P_{miss} (%)	P_{false} (%)	EER (%)
0.8959±0.0157	85.36±1.99	16.72±2.11	0.9733±0.0213	96.47±1.82	4.8	2.36	7.07±3.65

respectively, of the system based on SVM for the binary classification task. Performance scores (DCF_{opt}, P_{miss}, P_{false} and AUC) according to the average DET and ROC curve for one recording, are: $DCF_{opt} = 90.4\%$, $P_{miss} = 7.89\%$, $P_{false} = 10.18\%$ and $AUC = 0.9582$. These rates are worse than the average rates since they correspond to a single threshold T set for every run.

6.5 Discussion and Conclusions

Safe differentiation of innocent from abnormal heart murmurs in childhood is of paramount clinical significance [98]. Although the majority of children will manifest at some time an innocent

murmur, without any clinical significance, their discrimination from children having an abnormal murmur, due to an underlying heart malformation (only 1% prevalence) is a life saving clinical skill. However, clinical skills of physicians are rather declining as modern imaging modalities (echocardiography, MRI) allow for a detailed noninvasive diagnosis [74]. Nevertheless, offering a final diagnostic test as echocardiography to all children with a murmur although offers 100% diagnostic accuracy, cannot represent a viable solution, given the associated cost and personnel limitations [142]. Therefore, a heart disease screening system, especially if applied in low risk heart disease population as children should be still based on experienced cardiac auscultation, according to recent guidelines [92]. As primary health care physicians cardiac auscultation skills are rather suboptimal, experienced cardiac auscultation cannot be offered by them, if not having followed structured teaching programs in pediatric cardiac auscultation [45] (Erasmus Intensive Program "From Sound to Ultrasound: Multimedia based Pediatric Cardiology IP", 2010:www.med.uoc.gr/pcip2010). A computer based automated classification software, could be also of valuable help to the decision making of inexperienced physicians, based on the previous literature [111, 52, 127].

In the present study we aimed to develop and validate the clinical efficacy of an automatic detection and classification system of pediatric heart sounds based on reassigned spectrogram. Each phonocardiographic recording was automatically segmented into heart cycles by reference to the R-peaks of the ECG synchronous signal. We have evaluated features of the reassignment spectrogram of a fixed PCG segment for systolic murmur pathology detection. Different Time-Frequency (TF) distributions such as Exponential Time Distribution (ETD) and Hyperbolic Time Distribution (HTD) proposed in a previous study [109], have also been tested instead of the reassignment spectrogram. ETD and HTD offered equivalent resolution to the reassignment spectrogram [41]. However, a major drawback towards their use was the presence of several cross-terms depending on the choice of certain parameters. Therefore, the reassignment spectrogram was used as the initial time-frequency representation for the suggested classification system.

Features relevance to the task was assessed based on MI estimation. A great number of features was found necessary to perform discrimination. In order to extract a compact set of features out of this multidimensional representation, we removed "redundancy" at the first step of our processing, using HOSVD.

Our results show that appropriately selected reassignment spectral features are well suited to heart pathology assessment task. Using support vector machines (SVM) for systolic murmur classification, the suggested approach achieved an Equal Error Rate (EER) of $7.07 \pm 3.65\%$ and

an Area Under the Curve (AUC) score of 0.9733 ± 0.0213, which were comparable to the accuracy achieved by experienced pediatric cardiologists on the same database. This method could serve as a first step towards remote diagnosis and training of general physicians and pediatricians in accurate heart auscultation of children.

Previous studies have presented better efficiency scores for pathological systolic murmur discrimination using artificial neural networks [26], a thresholding technique based on the murmur intensity of a time-scale representation (continuous wavelet transform) [127], etc. However, we must emphasize that the data selection is crucial for the performance of a diagnostic system: the more clinically obvious the differences in the selected input dataset are regarding the variable studied (murmur), the better the performance of an automated diagnostic system will be (for example comparing patients without a murmur or hardly to detect innocent murmurs with those with abnormal murmurs so loud to be easily detected as such by inexperienced physicians). In the present study, we selected cases belonging rather to the gray area of easily misclassified systolic murmurs by inexperienced observers, being either loud innocent murmurs radiating to all sites, or abnormal murmurs associated with mild or moderate heart defects. It has been previously demonstrated that in the setting of mild heart defects even the performance of experienced pediatric cardiologists is rather suboptimal [44]. Normal subjects without any heart murmur were not included in the database we used in contrast to previous studies [109]. Furthermore, no specific recording sites corresponding to maximum murmur intensity were used; instead, we analyzed the signal of the first 3 recording positions which were available in each case. Hence, the automatic diagnostic system had to rely on characteristics of pathological heart sounds additional to the intensity values of the systolic murmurs: a wide split of the second tone (Figure 6.3a), or an early systolic click immediately after the first tone (Figure 6.3b).

False alarm (low specificity) of an automatic screening system, leads to the misclassification of innocent murmurs as pathological. Obviously, this is more tolerable compared to the misclassification of pathological murmurs as innocent (low sensitivity). System sensitivity could be further improved by employing additional temporal information such as the variability of the wide split of the second heart tone with respiration, or the variable intensity of the first heart sound between consequent heart cycles [43]. ECG and echocardiography are advisable then. In any case, the patient history and careful physical examination by a medical doctor are indispensable in clinical practice.

In conclusion, the performance of our automated approach was comparable to the reported accuracy achieved by experienced pediatric cardiologists and was significantly better than that

of general pediatricians [44]. Automated murmur classification methods could serve as a first line tool for clinical screening of heart disease in children, and a useful aid in teaching and practice of pediatric cardiac auscultation for non-cardiologists.

Chapter 7

General Conclusion

7.1 Contributions of this thesis

Time-frequency and modulation frequency analyses enable the classification of signals "with a considerably greater reflection of the physical situation that can be achieved by spectrum alone" ([67], p.26). The drawback is that features need to be extracted directly from the complete distribution, making necessary the use of advanced feature extraction and pattern recognition techniques. This thesis presented some advances in the field of feature selection from the multi-dimensional representations produced by the time-frequency or modulation-frequency analysis of audio signals. Solutions proposed were shown to provide appreciable improvements in various audio classification tasks. The original contributions of this thesis are the following:

Spectro-temporal Modulation Index Ratio

For the speech-nonspeech discrimination task based on direct pattern similarity in TIMIT, we adapted the spectro-temporal modulation index (STMI) - which is used to the assessment of speech intelligibility [34] - in order to handle the contribution of different frequency bands. Moreover, we took into account the similarity of audio signals to both cluster "prototypes" (speech and non-speech), by taking their ratio. We calculated the STMI and corresponding ratio (R) for all training examples and noise conditions. The histograms of STMI and R computed on speech and non-speech examples, formed two distinct clusters. A simple threshold check was used for discriminating speech from non-speech events. In the case of STMI, the overlap of the two "clusters" was small, but the decision threshold depended on the SNR condition - especially for low SNR (0dB, -10dB). In the case of R distribution, the overlap was slightly increased, however the

decision threshold was rather insensitive to the variation of SNR. This trend was reflected in the results in the benchmark test presented in [91].

Dimensionality and noise reduction of two-dimensional representations

As a first step for the dimensionality (and noise) reduction of modulation spectra and time-frequency distributions, we proposed a generalization of Singular Value Decomposition (SVD) algorithm to tensors (i.e., multi-dimensional matrices), referred to as Higher Order SVD. The singular value decomposition (SVD) of a time-frequency distribution was first proposed in [80] for the Wigner distribution. HOSVD of multiscale spectro-temporal modulations was first done in [96]. Comparing to SVD, reduction in dimensionality of feature space through HOSVD is performed in every subspace separately.

The $n-$rank (R_n) of a tensor $\mathbf{B}$ is the generalization of the column (row) rank of matrices: it equals the dimension of the vector space spanned by the n-mode vectors (the maximum number of linearly independent n-mode vectors); in contrast to the column and row rank of a matrix, the different n-ranks of a tensor are not necessarily the same. The truncation of the HOSVD maps each signal to a rank-(R_1, R_2) approximation. The appropriate values of R_1, R_2 can be determined by inspection of the singular value spectra in the respective modes. The ordering of the singular values in every mode implies that the "energy" of the tensor $\mathbf{B}$ is mainly concentrated in the part corresponding to low values of indices, i.e., the "significant" singular values. Real signals contain noise which is typically spread out over all the terms of the HOSVD decomposition. Since signals are well represented by the first few terms, truncating the series after the first few terms in every subspace, significantly reduces noise while retaining most of the signal [67].

HOSVD method seeks the projections along which the variance is higher, in order to represent the data in a low - dimensional space. Dimensions are discarded based on the relative magnitude of the corresponding singular values, without testing if these could be useful for classification. According to the maximum contribution (or minimum redundancy) criterion for feature selection, the optimal R_1, R_2 values for the rank-(R_1, R_2) approximation of audio signals, can be determined by cross-validation experiments for the classification. Still, this might be insufficient for selecting highly discriminative features; we employed a supervised learning algorithm towards this end.

Maximal Statistical Dependency Criterion and Most Relevant of Least Redundant features

The "m best features" are not necessarily the "best m features" due to the joint dependencies between features and target class [101]. In other words, when two features highly depend on each other, removing one of them from the feature set would not seriously affect its class-discriminative ability. The purpose of feature selection is to find a feature set S with m features $\{x_i\}$ which *jointly* have the largest dependency on the target class c; dependency of variables is usually defined based on mutual information [22]. This is the Max-Dependency scheme [101]. We have shown the Max-Dependency scheme to be equivalent to the "Maximal-Relevance" criterion when redundancy of features has been minimized before, for the "first-order" incremental search (i.e., one feature is added at one time during feature selection; cf Chapter 3). Through Maximal-Relevance scheme, the joint effects of features on the target class need not to be considered. In effect, we avoid the estimation of multivariate densities which is more difficult and less accurate than calculating bivariate densities.

High resolution energy representations can be highly redundant. Based on real data from MEEI database (sustained vowels) [37]), we have shown that HOSVD can significantly reduce the features redundancy, measured in terms of mutual information between features (cf Chapter 3). We have shown the superiority of our approach over minimum redundancy criterion based on cross-validation experiments for speech detection in Greek and U.S. English broadcast news (in Chapter 3) and voice pathology detection and classification (in Chapter 4).

Visualization through back-projection of mutual information in the original feature space

In Chapter 4 some interesting images were produced using the back-projection (eq. 3.6) of the most informative compact features to the original space; these "informative" images (Figure 4.4b) appear distorted relative to the rank$-(R_1, R_2)$ approximation of the signals resulting from the maximum contribution criterion. Still, the energy at modulations corresponding to pitch and syllabic and phonetic rates of speech (< 40 Hz) remain prominent.

Normalization of modulation spectra Using Mutual Information

In Chapter 5 we estimated the mutual information (MI) of modulation spectral features for the voice pathology detection task in two different databases (MEEI [37] and PdA [49]). We compared

different normalization schemes plus their combination, based on the assumption that the most relevant features selected in each database, should be similarly localized in the modulation-frequency plane. Subsequently we validated the normalization scheme through cross-database experiments on voice pathology detection task, i.e., training was performed on the first database (MEEI) whereas testing was done on the other (PdA); and vice versa. An interesting outcome of these experiments was that a more robust classification system was built when training was performed on the PdA database; the reason might be that in MEEI database normal speakers were recorded in a different environment than pathological speakers. Hence the most relevant features detected in MEEI (without any normalization), rather included channel specific characteristics.

Classification of Concatenated feature vectors vs fusion of classifiers on separate feature vectors

We have shown in Chapters 4 and 5 that the modulation spectral features are complementary to the state-of-the-art MFCC features for the speech detection and voice pathology detection tasks, respectively. Moreover, in the field of voice pathology detection, we compared the classification of concatenated feature vectors vs fusion of classifiers on separate feature vectors. Fusion of classifier scores can be an optimization problem by itself, since weights should be determined over the different classifiers outcome [65]. The two approaches were found to be rather equivalent, so we have chosen the first approach for its simplicity.

Classification of Mean of MFCC feature vectors vs Median of classification decisions

The minimum "optimum" time segment for the classification of speech signals has been shown to be comparable to the average length of a syllable in speech (250 ms) [96]. We have shown in Chapter 4 that classification of the mean of MFCC feature vectors over a time segment using SVM, is equivalent to taking the median of the classification decisions over the short-time feature vectors (corresponding to the same time segment). The reason for preferring the first approach is a significant reduction in the storage requirements of features, for training, validation and testing of a system, especially for large datasets.

Classification of fixed Systolic PCG Segments

Classification of systolic heart murmurs from phonocardiogram recordings (PCG), has been usually based on the proper alignment of the first two cardiac tones in every heart cycle and subsequent time-warping of the systolic interval. However, even if the first tone detection can be

accomplished with a 100% accuracy (based on the PCG-synchronous ECG and the mean spectral amplitude of the PCG over the frequencies $\nu \in 40-100$ Hz [32]), the detection of the second tone is prone to errors in the case of pathological murmurs [32]. Errors introduced in the detection of the second tone make difficult the proper temporal alignment of systolic heart segments (through a dynamic time-warping technique [35]). We have taken into account the relative stable duration of the systolic segments in every heart cycle as reported in [32]. Hence, we have performed classification of the time-frequency representation of fixed 400ms systolic segments from every heart cycle (cf Chapter 6). The direct comparison of qualitative characteristics (such as spectral patterns and time intervals in Figure 6.5) alleviates the need to detect and align the two cardiac tones in every heart cycle.

7.2 Perspectives

We have presented an approach towards classification of multi-dimensional representations resulting from time-frequency and modulation-frequency analysis of audio signals. More specifically, modulation spectral representations have been evaluated and compared on a large amount of data with the state-of-the-art techniques for the same classification task. In the case of systolic heart murmur classification, our approach was employed on reassigned spectra and compared to the paediatric cardiologists accuracy level on the same dataset.

The scope of these tools is not limited only to speech or audio signals; the same approach could be applied to the classification of general signals based on their time-frequency analysis. The choice of a particular time-frequency distribution still depends on the problem at hand.

The possible further investigations ensuing from the work presented in this thesis can be summarized as follows:

Voice Disorder Detection and Classification

Modulation spectral features have been shown to be complementary to the state-of-the-art MFCC features for automatic voice pathology detection (cf Chapter 5). Preliminary studies have also assessed their usefulness on the discrimination of pairs of various voice disorders, and voice quality assessment (grade and breathiness). Further work might include their assessment on connected speech (not only sustained vowels) and testing on bigger databases with a large number of recordings with different voice disorders.

Speaker recognition

A preliminary evaluation of the modulation spectral features for speaker recognition and validation has shown the speaker dependence of the significance of these features. A reason might be the voice quality variations (such as variation in the rates of glottalization) in normal speakers [112]. Given the complementarity between modulation spectra and MFCC features on speech and voice pathology detection tasks, it is reasonable to expect that combining them with a conventional speaker recognition system (e.g. with the typical GMM-MFCC approach) could lead to an improvement. Applying features normalization for channel compensation (as proposed in Chapter 5), could address the mismatch between training and testing conditions.

Rank properties of a tensor and SNR

The rank properties of the data tensor could be used towards determining the $I_n - R_n$ least significant singular values to be discarded in every mode, before feature selection proceeds. There is an analogy to the eigenvalue decomposition (EVD) of an observed covariance matrix $\mathbf{C_Y}$: in the case of spatially white noise, the noise covariance $\mathbf{C_E}$ is a diagonal matrix $\sigma_E^2\mathbf{I}$, in which σ_E^2 is the variance of the noise on each data channel (σ_E^{-1} is the signal-to-noise ratio, SNR); σ_E^2 can be estimated as the mean of the "noise-eigenvalues", i.e., the smallest $I - R$ eigenvalues of $\mathbf{C_Y}$ [68]. The $I_n - R_n$ least significant singular values in every mode then, could be related to the variance of the noise. We could determine the noise level as the mean of the least significant singular values in every subspace of the HOSVD decomposition.

On the other hand, if the noise levels (or SNR equivalently) are known, we could discard the singular values which are inferior or close enough to these levels. The least redundant features among which we select the most relevant ones, could be defined based on the minimum singular value which is superior to the noise level of the signals (SNR^{-1}).

It is worth noting that for high levels of noise (very low SNR), the significant truncation of the core tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, would mostly eliminate low variance components. In that case, our approach would not offer any advantage over minimum redundancy (maximum contribution) criterion combined with cross-validation (see next remark).

Heart murmur classification

In our last classification task, Max-Relevance first-order feature selection after HOSVD was not superior over minimum redundancy (maximum contribution) criterion, both combined with cross-

validation (cf Chapter 6). The high levels of noise in the phonocardiographic (PCG) data might be one reason, as we explain above. Still, both approaches achieve a comparable accuracy level to the paediatric cardiologists on the same dataset.

Different time-frequency or time-scale representations have to be evaluated and compared for the case of systolic heart murmur classification [109]. In addition, there are various useful features (like MFCC [109]) and decomposition techniques (e.g., adaptive time-frequency transform algorithms [133]) that produce a compact model of the signal which could be evaluated on this task and compared to our approach. Finally, there is the option to directly consider a sequence of at least three or five consequent heart cycles; this would permit us to examine the variance of the second cardiac tone with respiration, which is considered a significant clue towards heart pathology detection [31].

Principled feature selection

In all the classification tasks presented in this thesis, the relevance of information in the representation of an audio signal is assessed by Shannon's mutual information [22]. Mutual information estimation is based on the probability distribution of the data; thus, it is is a measure of relevance which is defined independently of the classifier and the evaluation metric. The selection of the most relevant among the least redundant features proceeds then using a traditional model selection method: cross-validation combined with a Support Vector Machine (SVM) classifier [60] and Equal Error Rate as a performance metric.

However, approaches to a principled feature selection suggest that relevancy definition should consider the learner used for classification and the metric used for evaluating the learner's performance (see [132, 2] and references therein). As the authors in [132] theoretically prove, "optimal feature selection is possible only for special cases; design of optimal feature selection algorithms is attainable only by constraining the application domain in terms of classifiers and loss functions used and tailoring the algorithms in those terms." It would be therefore interesting to investigate the potential of using algorithms that provably discover the optimal feature set for a given classification task, classifier and performance metric, in the applications presented in this thesis.

Appendix

Mutual Information estimation

The mutual information between two random variables x_i and x_j is defined in terms of their joint probability density function (pdf) $P_{ij}(x_i, x_j)$ and the marginal pdf's $P_i(x_i)$, $P_j(x_j)$. Mutual information (MI) $I[P_{ij}]$ is a natural measure of the inter-dependency between those variables:

$$I[P_{ij}] = \int dx_i \int dx_j P_{ij}(x_i, x_j) \log_2 \left[\frac{P_{ij}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \right] \tag{1}$$

MI quantifies how much information the value of one variable provides about the other; it is invariable to any invertible transformation of the individual variables [22].

It is well-known that MI estimation from observed data is non-trivial when (all or some of) the variables involved are continuous-valued. Estimating $I[P_{ij}]$ from a finite sample requires regularization of $P_{ij}(x_i, x_j)$. The simplest regularization is to define b discrete bins along each axis. We can make an adaptive quantization (variable bin length) so that the bins are equally populated and the coordinate invariance of the MI is preserved [123]. The precision of features quantization also affects the sample size dependence of MI estimates [22]. Entropies are systematically underestimated and mutual information is overestimated according to:

$$I_{est}(b, N) = I_\infty(b) + A(b)/N + C(b, N) \tag{2}$$

where I_∞ is the extrapolation to infinite sample size and the term $A(b)$ increases with b [123]. There is a critical value, b^*, beyond which the term $C(b, N)$ in (2) becomes important. The optimal b^* is defined according to a procedure described in [123]: when data are shuffled, mutual information should be near zero for a smaller number of bins ($b < b^*$) while it increases for more bins ($b > b^*$).

List of Publications

The following papers have been published in conference proceedings, journals or book chapters - except the last one which is to be submitted:

1. Markaki M., Wohlmayr M. and Stylianou Y., *Speech - Nonspeech Discrimination using the Information Bottleneck Method and Spectro-Temporal Modulation Index*, InterSpeech ICSLP, 2007.

2. Wohlmayr M., Markaki M., and Stylianou Y., *Speech - Nonspeech Discrimination based on Speech-relevant Spectrogram Modulations*, EUSIPCO, 2007.

3. Markaki M., Wohlmayr M. and Stylianou Y., *Extraction of Speech-Relevant Information from Modulation Spectrograms*, Progress in Nonlinear Speech Processing, Springer Berlin / Heidelberg, Springer, pp. 78 - 88, 2007.

4. Markaki M., Karpov A., Apostolopoulos E., Astrinaki M., Stylianou Y. and Ronzhin A., *A Hybrid System for Audio Segmentation and Speech-endpoint Detection of Broadcast News*, SPECOM 2007.

5. Markaki M. and Stylianou Y., *Discrimination of Speech from nonspeech in broadcast news based on modulation frequency features*, ISCA Tutorial and Research Workshop: Speech Analysis and Processing for Knowledge Discovery, 2008.

6. Markaki M. and Stylianou Y., *Dimensionality Reduction of Modulation Frequency Features for Speech Discrimination*, InterSpeech, 2008.

7. Markaki M., Holzapfel A. and Stylianou Y., *Singing Voice Detection using Modulation Frequency Features*, ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, 2008.

8. Markaki M. and Stylianou Y., *Evaluation of Modulation Frequency Features for Speaker Verification and Identification*, EUSIPCO, 2009.

9. Markaki M. and Stylianou Y., *Using Modulation Spectra for Voice Pathology Detection and Classification*, IEEE EMBC, 2009.

10. Markaki M. and Stylianou Y., *Normalized Modulation Spectral Features for Cross-Database Voice Pathology Detection*, InterSpeech, 2009.

11. Markaki M. and Stylianou Y., *Discrimination of Speech from Nonspeeech in Broadcast News Based on Modulation Frequency Features*, Speech Communication doi:10.1016/j.specom.2010.08.007, 2010.

12. Markaki M. and Stylianou Y., *Modulation Spectral Features for Objective Voice Quality Assessment: the Breathiness case*, MAVEBA, 2009.

13. Markaki M., Stylianou Y., Arias-Londono J.D. and Godino-Llorente J.I., *Dysphonia Detection based on Modulation Spectral Features and Cepstral Coefficients*, ICASSP, 2010.

14. Markaki M. and Stylianou Y., *Modulation Spectral Features for Objective Voice Quality Assessment*, IEEE ISCCSP, 2010.

15. Arias-Londono J.D., Godino-Llorente J.I., Markaki M. and Stylianou Y., *On combining information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for automatic detection of pathological voices*, Logopedics Phoniatrics Vocology, 2010.

16. Markaki M. and Stylianou Y., *Voice Pathology Detection and Discrimination Based on Modulation Spectral Features*, IEEE Transactions on Speech and Audio Processing, 2011.

17. Markaki M., Germanakis I. and Stylianou Y., *Automatic classification of Systolic Heart Murmurs Based on Reassigned Spectrogram*, to be submitted to IEEE Transactions on Biomedical Engineering.

Bibliography

[1] C. Ahlstrom, P. Hult, P. Rask, J-E. Karlsson, E. Nylander, U. Dahlström, and P. Ask. Feature extraction for systolic heart murmur classification. *Ann. Biomed. Eng.*, 34(11):1666–1677, 2006.

[2] C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X.D. Koutsoukos. Local causal and Markov Blanket induction for causal discovery and feature selection for classification: Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010.

[3] J. Alonso, J. de Leon, I. Alonso, and M. Ferrer. Automatic detection of pathologies in the voice by HOS based parameters. *Journal on Applied Signal Processing*, 4:275–284, 2001.

[4] J.D. Arias-Londono, J.I. Godino-Llorente, M. Markaki, and Y. Stylianou. On combining information from modulation spectra and Mel-frequency cepstral coefficients for automatic detection of pathological voices. *Logopedics Phoniatrics Vocology*, October 2010.

[5] H. Aronowitz. Segmental modeling for audio segmentation. In *ICASSP 2007*, pages 393–396, Hawaii, USA, 2006. IEEE.

[6] A. Askenfelt and B. Hammarberg. Speech waveform perturbation analysis revisited. *Speech Transmission Lab. - Quartely Progress and Status Report*, 22(4):49–68, 1981.

[7] L. Atlas and S. Schimmel. Modulation Toolbox for Matlab. http://isdl.ee.washington.edu/projects/modulationtoolbox/, 2005.

[8] L. Atlas and S.A. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 7:668–675, 2003.

[9] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Process.*, 43(5):1068–1089, 1995.

[10] F. Auger, P. Flandrin, P. Goncalves, and O. Lemoine. Time-frequency toolbox. http://tftb.nongnu.org, 2005.

[11] R.J. Baken. *Clinical measurement of speech and voice*. College Hill Press, Boston, 1987.

[12] H Barlow. *Possible principles underlying the transformation of sensory messages*, pages 217–234. MIT, Cambridge, MA, 1961.

[13] H. Barlow. Redundancy reduction revisited. *Network*, 12:241–253, 2001.

[14] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Trans. Audio, Speech and Language Proc.*, 14(5):1505–1512, 2006.

[15] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The curse of dimensionality for local kernel machines. Technical report, Department of Informatics and Operational Research, University of Montreal, 2005.

[16] S.R. Bhatikar, C. DeGroff, and R.L. Mahajan. A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. *Artificial Intelligence in Medicine*, 33:251–260, 2005.

[17] J.P.Jr. Campbell. Testing with the YOHO CD-ROM voice verification corpus. In *Proc. ICASSP*, volume 1, pages 341–344, 1995.

[18] RP Carlyon and SA Shamma. An account of monaural phase sensitivity. *J Acoust Soc Am*, 114(1):333–346, 2003.

[19] M.D. Cheitlin, J.S. Alpert, W.F. Armstrong, G.P. Aurigemma, G.A. Beller, F.Z. Bierman, T.W. Davidson, J.L. Davis, P.S. Douglas, L.D. Gillam, R.P. Lewis, A.S. Pearlman, J.T. Philbrick, P.M. Shah, R.G. Williams, J.L. Ritchie, K.A. Eagle, T.J. Gardner, A. Garson Jr, R.J. Gibbons, R.A. O'Rourke, and T.J. Ryan. ACC/AHA guidelines for the clinical application of echocardiography. A report of the American College of Cardiology/American Heart Association task force on practice guidelines (Committee on clinical application of echocardiography). *Circulation*, 95(6):1686–1744, 1997.

[20] T Chi, Y Gao, MC Guyton, P Ru, and S.A. Shamma. Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.*, 106:2719–2732, 1999.

[21] T Chi and SA Shamma. Spectrum restoration from multiscale auditory phase singularities by generalized projections. *IEEE Transactions on Speech and Audio Processing*, pages 1–14, 2006.

[22] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[23] L.B. Dahl, P. Hasvold, E. Arild, and T. Hasvold. Heart murmurs recorded by a sensor based electronic stethoscope and e-mailed for remote assessment. *Arch Disease Childhood*, 87(4):297–301, 2002.

[24] S.B. Davis. Computer evaluation of laryngeal pathology based on inverse filtering of speech. SCRL Monograph Number 13, 1976.

[25] S.M. Debbal and F. Bereksi-Reguig. Time-frequency analysis of the first and the second heartbeat sounds. *Applied Mathematics Computing*, 2(184):1041–1052, 2007.

[26] C.G. DeGroff, S. Bhatikar, J. Hertzberg, R. Shandas, L. Valdes-Cruz, and R.L. Mahajan. Artificial neural network-based method of screening heart murmurs in children. *Circulation*, 103:2711–2716, 2001.

[27] J.R. Deller, J.H.L. Hansen, and J. G. Proakis. *Discrete-time processing of speech signals*. McMillan, NY, 1993.

[28] A.A. Dibazar, T.W. Berger, and S.S. Narayanan. Pathological voice asessment. In *IEEE, 28th Eng. in Med. and Biol. Soc.*, pages 1669–1673, NY, NY, USA, August 2006.

[29] A.A. Dibazar and S.S. Narayanan. A system for automatic detection of pathological speech. In *36th Asilomar Conf. Signal, Systems, and Computers*, Asilomar, CA, USA, October 2002.

[30] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* Wiley-Interscience, New York, NY, 2000.

[31] M. El-Segaier. *Digital Analysis of Cardiac Acoustic signals in children.* Lunds University, Sweden, PhD Thesis, 2007.

[32] M. El-Segaier, O. Lilja, S. Lukkarinen, L. Sörnmo, R. Sepponen, and E. Pesonen. Computer-based detection and analysis of heart sound and murmur. *Ann. Biomed. Eng.*, 33(7):937–942, 2005.

[33] M. El-Segaier, E. Pesonen, S. Lukkarinen, K. Peters, L. Sörnmo, and R. Sepponen. Detection of cardiac pathology: time intervals and spectral analysis. *Acta Paediatrica*, 96(7):1036–1042, 2007.

[34] M Elhilali, T. Chi, and SA Shamma. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, 41:331–348, 2003.

[35] D. Ellis. Dynamic Time Warp in MATLAB. *http://www.ee.columbia.edu/ dpwe/resources/ matlab/dtw/*, 2003.

[36] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee. An information-theoretic perspective on feature selection in speaker recognition. *IEEE Signal Processing Letters*, 12:500–503, July 2005.

[37] Massachusetts Eye and Ear Infirmary. Elemetrics Disordered Voice Database (Version 1.03). Voice and Speech Lab, Boston, MA, October 1994. Kay Elemetrics Corp.

[38] C Ferencz, J.D. Rubin, R.J. McCarter, J.I. Brenner, C.A. Neill, L.W. Perry, S.I. Hepner, and J.W. Downing. Congenital heart disease: prevalence at livebirth: the Baltimore-Washington infant study. *Am J Epidem*, 121(1):31–36, 1985.

[39] J.P. Finley, A.E. Warren, G.P. Sharratt, and M. Amit. Assessing children's heart sounds at a distance with digital recordings. *Pediatrics*, 118(6):2322–2325, 2006.

[40] P. Flandrin. *Time-Frequency/Time-Scale Analysis.* Academic Press, San Diego, CA, 1999.

[41] P. Flandrin, F. Auger, and E. Chassande-Mottin. *Applications in Time-Frequency Signal Processing*, chapter Time-Frequency Reassignment - from principles to algorithms, pages 179–203. CRC Press, 2003.

[42] A. Fourcin and E. Abberton. Hearing and phonetic criteria in voice measurement: Clinical applications. *Logopedics Phoniatrics Vocology*, pages 1–14, April 2007.

[43] I. Germanakis. *Pediatric Cardiac Auscultation. Multimedia-based booklet.* I. Germanakis. ISBN: 978-960-93-2295-9, Athens, 2010.

[44] I. Germanakis, S. Dittrich, R. Perakaki, and M. Kalmanti. Digital Phonocardiography as a screening tool for heart disease in childhood. *Acta Paediatrica*, 51(3):327–333, 2008.

[45] I. Germanakis and M. Kalmanti. Pediatric cardiac auscultation teaching based on Digital Phonocardiography. *Medical Education*, 43(5):489, 2009.

[46] A. Giovanni, M. Ouaknine, and J.L. Triglia. Determination of largest Lyapunov exponents of vocal signal: Application to unilateral laryngeal paralysis. *Journal of Voice*, 13(3):341–354, 1999.

[47] J.I. Godino-Llorente and P. Gómez-Vilda. Automatic detection of voice impairments by means of short-time cepstral parameters and neural network-based detectors. *IEEE Trans. Biomed. Eng.*, 51(2):380–384, February 2004.

[48] J.I. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian Mixture Models and short-term cepstral parameters. *IEEE Trans. Biomed. Eng.*, 53(10):1943–1953, October 2006.

[49] J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo. Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program. *European Archives of Otolaringology*, 265(4):465–476, 2008.

[50] S. Greenberg and B.E.D. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *Proc. ICASSP*, volume 3, pages 1647–1650, 1997.

[51] S.W. Hainsworth and M.D. Macleod. Time Frequency Reassignment: A review and analysis. Cambridge University Engineering Department, 2003.

[52] C.S. Hayek and et al. Wavelet processing of systolic murmurs to assist with clinical diagnosis of heart disease. *Biomed Instrum Technol.*, 37(4):263–270, 2003.

[53] R.M. Hecht and N Tishby. Extraction of relevant speech features using the information bottleneck method. In *Proceedings of Interspeech, Lisbon*, 2005.

[54] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J.A.S.A.*, 87(4):1738–1752, 1990.

[55] H. Hermansky. Should recognizers have ears? *Speech Communication*, 25:3–27, August 1998.

[56] M. Hirano. Objective evaluation of the human voice: clinical aspects. *Folia Phoniatr.*, 41:89–144, 1989.

[57] V.K. Holldack and D. Wolf. *Phonokardiographie und verwandter Undersuchngsmethoden.* Georg Thieme Verlag, Stuttgart, 1974.

[58] P.T. Hosseini, F. Almasganj, T. Emami, R. Behroozmand, S. Gharibrade, and F. Torabinezhad. Local Discriminant Wavelet Packet Basis for voice pathology classification. In *2nd Intern. Conf. on Bioinformatics and Biomedial Eng. (ICBBE)*, pages 2052–2055, May 2008.

[59] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38:2270–2285, 2005.

[60] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-scale SVM Learning Practical. MIT-Press, 1999.

[61] M. Joos. Acoustic phonetics. *Lang. Monogr.*, 23:1–137, 1948.

[62] T. Kinunnen. Joint acoustic-modulation frequency for speaker recognition. In *Proc. ICASSP*, volume 1, pages 665–668. IEEE, 2006.

[63] T. Kinunnen, E. Chernenko, M. Tuononen, P. Franti, and H. Li. Voice activity detection using MFCC features and Support Vector Machine. In *Proc. SPECOM 2007*, volume 2, pages 556–561, 2007.

[64] T. Kinunnen, K.A. Lee, and H. Li. Dimension reduction of the modulation spectrogram for speaker verification. In *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008.

[65] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. and Machine Intel.*, 20(3):226–239, 1998.

[66] D. Kumar, P. Carvalho, M. Antunes, and J. Henriques. Noise detection during heart sound recording. In *IEEE EMBC09*, pages 3119–3123, 2009.

[67] Cohen L. *Time-Frequency Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1995.

[68] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[69] M.A. Little, P.E. McSharry, S.J. Roberts, D. Costello, and I.M.Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering, Published online, doi:10.1186/1475-925X-6-23*, June 2007.

[70] L. Lu, H.J. Zhang, and S. Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems*, 8:482–492, 2003.

[71] C. Mahnke. Automated heartsound analysis/computer-aided ausculation: A cardiologist's perspective and suggestions for future development. In *IEEE EMBC09*, 2009.

[72] N. Malyska, T.F. Quatieri, and D. Sturim. Automatic dysphonia recognition using biologically inspired amplitude-modulation features. In *Proc. ICASSP*, pages 873–876, 2005.

[73] S. Mangione. Cardiac auscultatory skills of physicians-in-training: a comparison of three english-speaking countries. *Am J Med*, 110:210–216, 2001.

[74] S. Mangione and L.Z. Nieman. Cardiac auscultatory skills of internal medicine and family practice trainees. a comparison of diagnostic proficiency. *JAMA*, 278:717–722, 1997.

[75] D. Manoussaki, R.S. Chadwick, D.R. Ketten, J. Arruda, E. Dimitriadis, and J.T. O'Malley. The influence of cochlear shape on low-frequency hearing. *Proc Natl Acad Sci USA*, 105(16):6162–6166, 2008.

[76] D. Manoussaki, E. Dimitriadis, and R.S. Chadwick. The cochlea's graded curvature effect on low frequency waves. *Physical Review Letters*, 96(8), 2006.

[77] P. Maragos, J.F. Kaiser, and T.F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. on Signal Processing*, 41(10):3024–3051, 1993.

[78] P. Maragos, J.F. Kaiser, and T.F. Quatieri. On amplitude and frequency demodulation using energy operators. *IEEE Trans. on Signal Processing*, 41(4):1532–1550, 1993.

[79] P. Maragos, T.F. Quatieri, and J.F. Kaiser. Speech nonlinearities, modulations and energy operators. In *ICASSP 1991*, pages 421–424, Toronto, Canada, 1991. IEEE.

[80] N.M. Marinovich and G. Eichmann. An expansion of the Wigner distribution and its applications. In *Proc. ICASSP*, volume 3, pages 1021–1024. IEEE, 1985.

[81] M. Markaki and Y. Stylianou. Dimensionality reduction of modulation frequency features for speech discrimination. In *Proc. Interspeech 2008*, pages 646–649, Brisbane, Australia, 2008.

[82] M. Markaki and Y. Stylianou. Discrimination of speech from nonspeech in broadcast news based on modulation frequency features. In *ISCA Tutorial and Research Workshop (ITRW) on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, 2008.

[83] M. Markaki and Y. Stylianou. Singing voice detection using modulation frequency features. In *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audition*, Brisbane, Australia, 2008.

[84] M. Markaki and Y. Stylianou. Evaluation of modulation frequency features for speaker verification and identification. In *Proc. EUSIPCO 2009*, Glasgow, Scotland, 2009.

[85] M. Markaki and Y. Stylianou. Modulation spectral features for objective voice pathology assessment: the breathiness case. In 6^{th} *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Florence, Italy, 2009.

[86] M. Markaki and Y. Stylianou. Normalized modulation spectral features for cross-database voice pathology detection. In *Proc. Interspeech*, Brighton, U.K., 2009.

[87] M. Markaki and Y. Stylianou. Using modulation spectra for voice pathology detection and classification. In *Proceedings of IEEE EMBC'09*, Minneapolis, Minnesota, U.S.A., 2009.

[88] M. Markaki and Y. Stylianou. Modulation spectral features for objective voice quality assessment. In *Proc. IEEE ISCCSP*, Limassol, Cyprus, 2010.

[89] M. Markaki and Y. Stylianou. Voice pathology detection and discrimination based on modulation spectral features. *IEEE Trans. Speech Audio Process.*, 2011.

[90] M. Markaki, Y. Stylianou, J.D. Arias-Londono, and J.I. Godino-Llorente. Dysphonia detection based on modulation spectral features and cepstral coefficients. In *Proc. IEEE ICASSP*, 2010.

[91] M. Markaki, M. Wohlmayr, and Y. Stylianou. Speech - nonspeech discrimination using the Information Bottleneck method and Spectro-Temporal Modulation Index. In *Proc. Interspeech-ICSLP 2007*, Antwerp, Belgium, 2007.

[92] B.J. Maron, P.D. Thompson, M.J. Ackerman, G. Balady, S. Berger, D. Cohen, R. Dimeff, P.S. Douglas, D.W. Glover, A.M. Hutter Jr, M.D. Krauss, M.S. Maron, M.J. Mitten, W.O. Roberts, and J.C. Puffer. Recommendations and Considerations Related to Preparticipation Screening for Cardiovascular Abnormalities in Competitive Athletes: 2007 Update: A Scientific Statement From the American Heart Association Council on Nutrition, Physical Activity, and Metabolism: Endorsed by the American College of Cardiology Foundation. *Circulation*, 115(12):1643–1655, 2007.

[93] S.L. Marple. *Digital spectral analysis with applications*. Prentice-Hall, NJ, 1987.

[94] A. Martin, G.R. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, volume IV, pages 1895–1898, 1997.

[95] N Mesgarani, M Slaney, and SA Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Speech and Audio Processing*, PP(99):1–11, 2006.

[96] N. Mesgarani, M. Slaney, and S.A. Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Trans. Speech Audio Process.*, 14:920–930, 2006.

[97] I Nelken, Y Rotman, and O Bar Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397:154–7, 1999.

[98] J.W. Newburger, A. Rosenthal, R.G. Williams, K. Fellows, and O.S. Miettinen. Noninvasive tests in the initial evaluation of heart murmurs in children. *N Engl J Med*, 308:61–64, 1983.

[99] A.-L. Noponen, S. Lukkarinen, A. Angerla, and R. Sepponen. Phono-spectrographic analysis of heart murmur in children. *BMC Pediatrics*, 7(1):23, 2007.

[100] V. Parsa and D.G. Jamieson. Identification of pathological voices using glottal noise measures. *J. Speech, Language, Hearing Res.*, 43(2):469–485, April 2000.

[101] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1226–1238, 2005.

[102] A. Potamianos and P. Maragos. Speech analysis and synthesis using an AM-FM modulation model. *Speech Communication*, 28:195–209, 1999.

[103] G. Pouchoulin, C. Fredouille, J.F. Bonastre, A. Ghio, and A. Giovanni. Frequency study for the characterization of the dysphonic voices. In *Proc. Eurospeech*, pages 1198–1201, Antwerp, Belgium, 2007.

[104] R.A. Prosek, A.A. Montgomery, B.E. Walden, and D.B. Hawkins. An evaluation of residue features as correlates of voice disorders. *Journal of Communication Disorders*, 20:105–117, 1987.

[105] A. Qiu, C.E. Schreiner, and M.A. Escabi•. Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. *J Neurophysiol*, 90:456–476, 2003.

[106] M.D. Plumbe T.F. Quatieri and D.A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech Audio Process.*, 7:569–587, 1999.

[107] T.F. Quatieri. *Discrete Time Speech Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 2002.

[108] T.F. Quatieri, N. Malyska, and D.E. Sturim. Auditory signal processing as a basis for speaker recognition. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain, NY, 2003.

[109] A. F. Quiceno-Manrique, J.I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez. Selection of dynamic features based on time-frequency representations for heart murmur detection from phonocardiographic signals. *Ann Biomed Eng*, 38(1):118–137, 2010.

[110] K. Rajakumar, M. Weisse, A. Rosas, E. Gunel, L. Pyles, W.A. Neal, A. Balian, and S. Einzig. Comparative study of clinical evaluation of heart murmurs by general pediatricians and pediatric cardiologists. *Clinical Pediatrics*, 38(9):511–518, 1999.

[111] P. Ramnarayan and J. Britto. Paediatric clinical decision support systems. *Arch Disease Childhood*, 87(5):361–362, 2002.

[112] L. Redi and S. Shattuck-Hufnagel. Variation in the realization of glottalization in normal speakers. *J. Phonetics*, 29:407–429, 2001.

[113] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted Gaussian Mixture Models. *Digit. Signal Processing*, 10(1):19–41, 2000.

[114] O. Rioul and M. Vetterli. Wavelets and Signal Processing. *IEEE Signal Processing Magazine*, 8(4):14–38, 1991.

[115] M. Rocamora and P. Herrera. Comparing audio descriptors for singing voice detection in music audio files. In *Proc. 11th Brazilian Symposium on Computer Music*, San Pablo, Brazil, 2007.

[116] M. Rosa, J.C.Pereira, and M.Grellet. Adaptive estimation of residue signal for voice pathology diagnosis. *IEEE Trans. Biomed. Eng.*, 47(1):96–104, Jan 2000.

[117] C. Sanderson and K.K. Paliwal. Information fusion and person verification using speech and face information. Research Paper IDIAP-RR 02-33, 2002.

[118] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proc. ICASSP*, pages 993–996. IEEE, 1996.

[119] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature music/speech discriminator. In *Proc. ICASSP*, pages 1331–1334. IEEE, 1997.

[120] S.M. Schimmel, L.E. Atlas, and K. Nie. Feasibility of single channel speaker separation based on modulation frequency analysis. In *Proc. ICASSP*, volume 4, pages 605–608, 2007.

[121] N.C. Singh and F.E. Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.*, 114(6):3394–3411, Dec 2003.

[122] N. Slonim. The information bottleneck: Theory and applications. *School of Engineering and Computer Science*, 2002.

[123] N. Slonim, G.S. Atwal, G. Tkacik, and W. Bialek. Estimating mutual information and multi-information in large networks. http://arxiv.org/abs/cs.IT/0502017, 2005.

[124] Raimund Specht. Animal sound recordings, Avisoft Bioacoustics, 2006. www.avisoft.com.

[125] M. S. Spina and V. Zue. Automatic transcription of general audio data: Preliminary analyses. In *Proc. ICSLP '96*, volume 2, pages 594–597, Philadelphia, PA, 1996.

[126] S. Sukittanon, L. Atlas, and J.W. Pitton. Modulation-scale analysis for content identification. *IEEE Trans. Speech Audio Process.*, 52(10):3023–3035, 2004.

[127] W.R Thomson, C.S. Hayek, C. Tuchinda, J.K. Telford, and J.S. Lombardo. Automated cardiac auscultation for detection of pathologic heart murmurs. *Pediatric Cardiology*, 22:373–379, 2001.

[128] N. Tishby, F. Pereira, and W. Bialek. The Information Bottleneck method. *The 37th annual Allerton Conference on Communication, Control and Computing*, 1999.

[129] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[130] A. Tlkian and M. Conover. *Understanding Heart Sounds and Murmurs: With an Introduction to Lung Sounds*. W. B. Saunders Co., Philadelphia, 2001.

[131] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Trans. Audio, Speech and Language Proc.*, 14(5):1557–1565, 2006.

[132] I. Tsamardinos and C.F. Aliferis. Towards principled feature selection: relevancy, filters and wrappers. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 376–381, 2003.

[133] K. Umapathy, S. Krishnan, V. Parsa, and D.G. Jamieson. Discrimination of pathological voices using time-frequency approach. *IEEE Trans. Biomed. Eng.*, 52(3):421–430, 2005.

[134] M.N. Vieira, F.R. MClnnes, and M.A. Jack. On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures. *J.A.S.A.*, 111(2):1045–1055, 2002.

[135] G. von Bèkèsy. Travelling waves as frequency analysers in the cochlea. *Nature*, 225 (5239):1207–1209, 1970.

[136] V. Wan and S. Renals. Speaker verification using sequence discriminant support vector machines. *IEEE Trans. Audio, Speech and Language Proc.*, 13(2):203–210, 2005.

[137] K. Wang and S.A. Shamma. Spectral shape analysis in the central auditory system. *IEEE Transactions on Speech and Audio Processing*, 3(5):382–396, 1995.

[138] M. Wohlmayr, M. Markaki, and Y. Stylianou. Speech - nonspeech discrimination based on speech-relevant spectrogram modulations. In *Proc. EUSIPCO 2007*, Poznan, Poland, 2007.

[139] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: the ICSI-SRI Fall 2004 diarization system. In *Proceedings of Fall 2004 Rich Transcription Workshop*, 2004.

[140] H. Yang, S. van Vuuren, and H. Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *ICASSP Proceedings*, pages 3–27, 1999.

[141] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38(2):824–839, 1992.

[142] M.S. Yi, T.R. Kimball, J. Tsevat, J.M. Mrus, and U.R. Kotagal. Evaluation of heart murmurs in children: cost-effectiveness and practical implications. *J Pediatr*, 141:504–511, 2002.

[143] F-G Zeng, K Nie, G. S. Stickney, Y-Y Kong, M Vongphoe, A Bhargave, C Wei, and K Cao. Speech recognition with amplitude and frequency modulations. *Proc Natl Acad Sci USA*, 102(7):2293–2298, 2005.

[144] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J.J. Jiang. Structure and function of auditory cortex: music and speech. *Trends in Cognitive Neurosciences*, 6(1):37–46, 2002.

[145] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J.J. Jiang. Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps. *Journal of the Acoustical Society of America*, 115(5):2270–2277, 2004.