

## Audio Signal Processing

Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay,  
India [prao@ee.iitb.ac.in](mailto:prao@ee.iitb.ac.in)

### 1 Introduction

Our sense of hearing provides us rich information about our environment with respect to the locations and characteristics of sound producing objects. For example, we can effortlessly assimilate the sounds of birds twittering outside the window and traffic moving in the distance while following the lyrics of a song over the radio sung with multi-instrument accompaniment. The human auditory system is able to process the complex sound mixture reaching our ears and form high-level abstractions of the environment by the analysis and grouping of measured sensory inputs. The process of achieving the segregation and identification of sources from the received composite acoustic signal is known as auditory scene analysis. It is easy to imagine that the machine realization of this functionality (sound source separation and classification) would be very useful in applications such as speech recognition in noise, automatic music transcription and multimedia data search and retrieval. In all cases the audio signal must be processed based on signal models, which may be drawn from sound production as well as sound perception and cognition. While production models are an integral part of speech processing systems, general audio processing is still limited to rather basic signal models due to the diverse and wide-ranging nature of audio signals.

Important technological applications of digital audio signal processing are audio data compression, synthesis of audio effects and audio classification. While audio compression has been the most prominent application of digital audio processing in the recent past, the burgeoning importance of multimedia content management is seeing growing applications of signal processing in audio segmentation and classification. Audio classification is a part of the larger problem of audiovisual data handling with important applications in digital libraries, professional media production, education, entertainment and surveillance. Speech and speaker recognition can be considered classic problems in audio retrieval and have received decades of research attention. On the other hand, the rapidly growing archives of digital music on the internet are

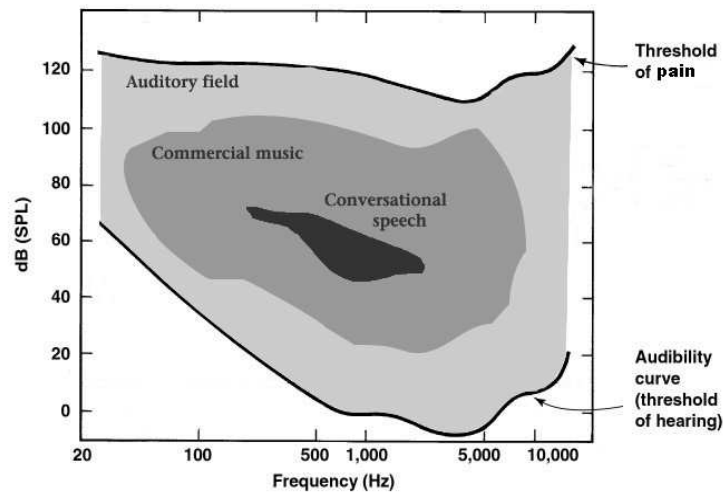
now drawing attention to wider problems of nonlinear browsing and retrieval using more natural ways of interacting with multimedia data including, most prominently, music. Since audio records (unlike images) can be listened to only sequentially, good indexing is valuable for effective retrieval. Listening to audio clips can actually help to navigate audiovisual material more easily than the viewing of video scenes. Audio classification is also useful as a front end to audio compression systems where the efficiency of coding and transmission is facilitated by matching the compression method to the audio type, as for example, speech or music.

In this chapter, we review the basic methods for signal processing of audio, mainly from the point of view of audio classification. General properties of audio signals are discussed followed by a description of time-frequency representations for audio. Features useful for classification are reviewed followed by a discussion of prominent examples of audio classification systems with particular emphasis on feature extraction.

## 2 Audio Signal Characteristics

Audible sound arises from pressure variations in the air falling on the ear drum. The human auditory system is responsive to sounds in the frequency range of 20 Hz to 20 kHz as long as the intensity lies above the frequency dependent “threshold of hearing”. The audible intensity range is approximately 120 dB which represents the range between the rustle of leaves and boom of an aircraft take-off. Figure 1 displays the human auditory field in the frequency-intensity plane. The sound captured by a microphone is a time waveform of the air pressure variation at the location of the microphone in the sound field. A digital audio signal is obtained by the suitable sampling and quantization of the electrical output of the microphone. Although any sampling frequency above 40 kHz would be adequate to capture the full range of audible frequencies, a widely used sampling rate is 44,100 Hz, which arose from the historical need to synchronize audio with video data. “CD quality” refers to 44.1 kHz sampled audio digitized to 16-bit word length.

Sound signals can be very broadly categorized into environmental sounds, artificial sounds, speech and music. A large class of interesting sounds is time-varying in nature with information coded in the form of temporal sequences of atomic sound events. For example, speech can be viewed as a sequence of phones, and music as the evolving pattern of notes. An atomic sound event, or a single gestalt, can be a complex acoustical signal described by a specific set of temporal and spectral properties. Examples of atomic sound events include short sounds such as a door slam, and longer uniform texture sounds such as the constant patter of rain. The temporal properties of an audio event refer to the duration of the sound and any amplitude modulations including the rise and fall of the waveform amplitude envelope. The spectral properties of the sound relate to its frequency components and their relative strengths.



**Fig. 1.** The auditory field in the frequency-intensity plane. The sound pressure level is measured in dB with respect to the standard reference pressure level of 20 microPascals.

Audio waveforms can be periodic or aperiodic. Except for the simple sinusoid, periodic audio waveforms are complex tones comprising of a fundamental frequency and a series of overtones or multiples of the fundamental frequency. The relative amplitudes and phases of the frequency components influence the sound “colour” or timbre. Aperiodic waveforms, on the other hand, can be made up of non-harmonically related sine tones or frequency shaped noise. In general, a sound can exhibit both tone-like and noise-like spectral properties and these influence its perceived quality. Speech is characterized by alternations of tonal and noisy regions with tone durations corresponding to vowel segments occurring at a more regular syllabic rate. Music, on the other hand, being a melodic sequence of notes is highly tonal for the most part with both fundamental frequency and duration varying over a wide range.

Sound signals are basically physical stimuli that are processed by the auditory system to evoke psychological sensations in the brain. It is appropriate that the salient acoustical properties of a sound be the ones that are important to the human perception and recognition of the sound. Hearing perception has been studied since 1870, the time of Helmholtz. Sounds are described in terms of the perceptual attributes of pitch, loudness, subjective duration and timbre. The human auditory system is known to carry out the frequency analysis of sounds to feed the higher level cognitive functions. Each of the subjective sensations is correlated with more than one spectral property (e.g. tonal content) or temporal property (e.g. attack of a note struck on an instrument) of the sound. Since both spectral and temporal properties are relevant

to the perception and cognition of sound, it is only appropriate to consider the representation of audio signals in terms of a joint description in time and frequency.

While audio signals are non stationary by nature, audio signal analysis usually assumes that the signal properties change relatively slowly with time. Signal parameters, or features, are estimated from the analysis of short windowed segments of the signal, and the analysis is repeated at uniformly spaced intervals of time. The parameters so estimated generally represent the signal characteristics corresponding to the time center of the windowed segment. This method of estimating the parameters of a time-varying signal is known as “short-time analysis” and the parameters so obtained are referred to as the “short-time” parameters. Signal parameters may relate to an underlying signal model. Speech signals, for example, are approximated by the well-known source-filter model of speech production. The source-filter model is also applicable to the sound production mechanism of certain musical instruments where the source refers to a vibrating object, such as a string, and the filter to the resonating body of the instrument. Music due to its wide definition, however, is more generally modelled based on observed signal characteristics as the sum of elementary components such as continuous sinusoidal tracks, transients and noise.

### 3 Audio Signal Representations

The acoustic properties of sound events can be visualized in a time-frequency “image” of the acoustic signal so much so that the contributing sources can often be separated by applying gestalt grouping rules in the visual domain. Human auditory perception starts with the frequency analysis of the sound in the cochlea. The time-frequency representation of sound is therefore a natural starting point for machine-based segmentation and classification. In this section we review two important audio signal representations that help to visualize the spectro-temporal properties of sound, the spectrogram and an auditory representation. While the former is based on adapting the Fourier transform to time-varying signal analysis, the latter incorporates the knowledge of hearing perception to emphasize perceptually salient characteristics of the signal.

#### 3.1 Spectrogram

The spectral analysis of an acoustical signal is obtained by its Fourier transform which produces a pair of real-valued functions of frequency, called the amplitude (or magnitude) spectrum and the phase spectrum. To track the time-varying characteristics of the signal, Fourier transform spectra of overlapping windowed segments are computed at short successive intervals. Time-domain waveforms of real world signals perceived as similar sounding actually

show a lot of variability due to the variable phase relations between frequency components. The short-time phase spectrum is not considered as perceptually significant as the corresponding magnitude or power spectrum and is omitted in the signal representation [1]. From the running magnitude spectra, a graphic display of the time-frequency content of the signal, or spectrogram, is produced.

Figure 1 shows the waveform of a typical music signal comprised of several distinct acoustical events as listed in Table 1. We note that some of the events overlap in time. The waveform gives an indication of the onset and the rate of decay of the amplitude envelope of the non-overlapping events. The spectrogram (computed with a 40 ms analysis window at intervals of 10 ms) provides a far more informative view of the signal. We observe uniformly-spaced horizontal dark stripes indicative of the steady harmonic components of the piano notes. The frequency spacing of the harmonics is consistent with the relative pitches of the three piano notes. The piano notes' higher harmonics are seen to decay fast while the low harmonics are more persistent even as the overall amplitude envelope decays. The percussive (low tom and cymbal crash) sounds are marked by a grainy and scattered spectral structure with a few weak inharmonic tones. The initial duration of the first piano strike is dominated by high frequency spectral content from the preceding cymbal crash as it decays. In the final portion of the spectrogram, we can now clearly detect the simultaneous presence of piano note and percussion sequence.

**Table 1.** A description of the audio events corresponding to Figure 1.

Time duration (seconds)	Nature of sound event
0.00 to 0.15	Low Tom (Percussive stroke)
0.15 to 1.4	Cymbal crash (percussive stroke)
1.4 to 2.4	Piano note (Low pitch)
2.4 to 3.4	Piano note (High pitch)
3.4 to 4.5	Piano note (Middle pitch) occurring simultaneously as the low tom and cymbal crash (from 0 to 1.4 sec)

The spectrogram by means of its time-frequency analysis displays the spectro-temporal properties of acoustic events that may overlap in time and frequency. The choice of the analysis window duration dictates the trade-off between the frequency resolution of steady-state content versus the time resolution of rapidly time-varying events or transients.

### 3.2 Auditory Representations

Auditory representations attempt to describe the acoustic signal by capturing the most important aspects of the way the sound is perceived by humans.

Compared with the spectrogram, the perceptually salient aspects of the audio signal are more directly evident. For instance, the spectrogram displays the time-frequency components according to their physical intensity levels while the human ear’s sensitivity to the different components is actually influenced by several auditory phenomena. Prominent among these are the widely differing sensitivities of hearing in the low, mid and high frequency regions in terms of the threshold of audibility (see Figure 1), the nonlinear scaling of the perceived loudness with the physically measured signal intensity and the decreasing frequency resolution, with increasing frequency, across the audible range [2]. Further, due to a phenomenon known as auditory masking, strong signal components suppress the audibility of relatively weak signal components in their time-frequency vicinity. The various auditory phenomena are explained based on the knowledge of ear physiology combined with perception models derived from psychoacoustical listening experiments [2].

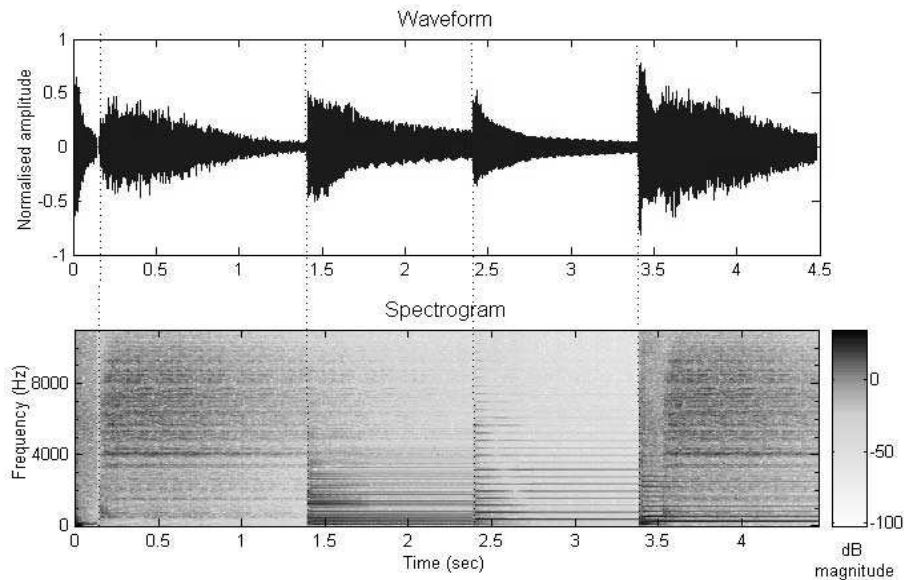
An auditory representation is typically obtained as the output of a computational auditory model and represents a physical quantity at some stage of the auditory pathway. Computational auditory models simulate the outer, middle and inner ear functions of transforming acoustic energy into a neural code in the auditory nerve. The computational models are thus based on the approximate stages of auditory processing with the model parameters fitted to explain empirical data from psychoacoustical experiments. For instance, cochlear models simulate the band-pass filtering action of the basilar membrane and the subsequent firing activity of the hair-cell neurons as a function of place along the cochlea. Several distinct computational models have been proposed over the years [3], [4]. However a common and prominent functional block of the cochlear models is the integration of sound intensity over a finite frequency region by each of a bank of overlapping, band-pass filters. Each of these “critical bandwidth” filters corresponds to a cochlear channel whose output is processed at the higher levels of the auditory pathway relatively independently of other channels. The filter bandwidths increase roughly logarithmically with center frequency indicating the non-uniform frequency resolution of hearing. The output of the cochlear model is a kind of “auditory image” that forms the basis for cognitive feature extraction by the brain. This is supported by the psychoacoustic observation that the subjective sensations that we experience from the interaction of spectral components that fall within a cochlear channel are distinctly different from the interaction of components falling in separate channels [2].

In the Auditory Image Model of Patterson [3], the multichannel output of the cochlea is simulated by a gammatone filterbank followed by a bank of adaptive thresholding units simulating neural transduction. Fourth-order gammatone filters have frequency responses that can closely approximate psychoacoustically measured auditory filter shapes. The filter bandwidths follow the measured critical bandwidths ( $25 \text{ Hz} + 10\%$  of the center frequency) and center frequencies distributed along the frequency axis in proportion to their bandwidths. The filter outputs are subjected to rectification, compression and

low-pass filtering before the adaptive thresholding. The generated auditory image is perceptually more accurate than the spectrogram in terms of the time and frequency resolutions of signal components and their perceived loudness levels.

## 4 Audio Features for Classification

While the spectrogram and auditory signal representations discussed in the previous section are good for visualization of audio content, they have a high dimensionality which makes them unsuitable for direct application to classification. Ideally, we would like to extract low-dimensional features from these representations (or even directly from the acoustical signal) which retain only the important distinctive characteristics of the intended audio classes. Reduced-dimension, decorrelated spectral vectors obtained using a linear transformation of a spectrogram have been proposed in MPEG-7, the audiovisual content description standard [5], [6].



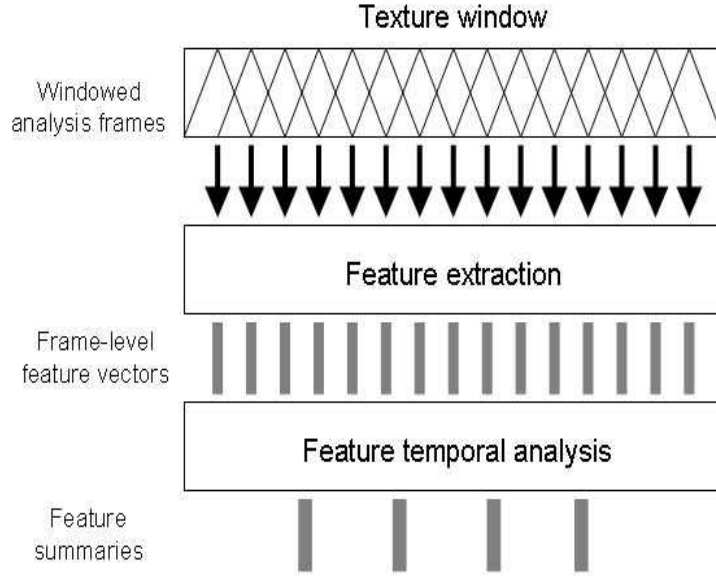
**Fig. 2.** (a) Waveform, and (b) spectrogram of the audio segment described in Table 1. The vertical dotted lines indicate the starting instants of new events. The spectrogram relative intensity scale appears at lower right.

A more popular approach to feature design is to use explicit knowledge about the salient signal characteristics either in terms of signal production or

perception. The goal is to find features that are invariant to irrelevant transformations and have good discriminative power across the classes. Feature extraction, an important signal processing task, is the process of computing the numerical representation from the acoustical signal that can be used to characterize the audio segment. Classification algorithms typically use labeled training examples to partition the feature space into regions so that feature vectors falling in the same region come from the same class. A well-designed set of features for a given audio categorization task would make for robust classification with reasonable amounts of training data. Audio signal classification is a subset of the larger problem of auditory scene analysis. When the audio stream contains many different, but non-simultaneous, events from different classes, segmentation of the stream to separate class-specific events can be achieved by observing the transitions in feature values as expected at segment boundaries. However when signals from the different classes (sources) overlap in time, stream segregation is a considerably more difficult task [7].

Research on audio classification over the years has given rise to a rich library of computational features which may be broadly categorized into physical features and perceptual features. Physical features are directly related to the measurable properties of the acoustical signal and are not linked with human perception. Perceptual features, on the other hand, relate to the subjective perception of the sound, and therefore must be computed using auditory models. Features may further be classified as static or dynamic features. Static features provide a snapshot of the characteristics of the audio signal at an instant in time as obtained from a short-time analysis of a data segment. The longer-term temporal variation of the static features is represented by the dynamic features and provides for improved classification. Figure 3 shows the structure of such a feature extraction framework [8]. At the lowest level are the analysis frames, each representing windowed data of typical duration 10 ms to 40 ms. The windows overlap so that frame durations can be significantly smaller, usually corresponding to a frame rate of 100 frames per second. Each audio frame is processed to obtain one or more static features. The features may be a homogenous set, like spectral components, or a more heterogenous set. That is, the frame-level feature vector corresponds to a set of features extracted from a single windowed audio segment centered at the frame instant. Next the temporal evolution of frame-level features is observed across a larger segment known as a texture window to extract suitable dynamic features or feature summaries. It has been shown that the grouping of frames to form a texture window improves classification due to the availability of important statistical variation information [9], [10]. However increasing the texture window length beyond 1 sec does not improve classification any further. Texture window durations typically range from 500 ms to 1 sec. This implies a latency or delay of up to 1 sec in the audio classification task.





**Fig. 3.** Audio feature extraction procedure (adapted from [8]).

#### 4.1 Physical Features

Physical features are low-level signal parameters that capture particular aspects of the temporal or spectral properties of the signal. Although some of the features are perceptually motivated, we classify them as physical features since they are computed directly from the audio waveform amplitudes or the corresponding short-time spectral values. Widely applied physical features are discussed next. In the following equations, the subindex “r” indicates the current frame so that  $x_r[n]$  are the samples of the N-length data segment (possibly multiplied by a window function) corresponding to the current frame. We then have for the analysis of the  $r^{th}$  frame,

$$\left\{ \begin{array}{c} x_r[n] \\ n = 1 \dots N \end{array} \right\} \longrightarrow \left\{ \begin{array}{c} X_r[k] \text{ at freq. } f[k] \\ k = 1 \dots N \end{array} \right\} \quad (1)$$

#### Zero-Crossing Rate

The Zero-Crossing Rate (ZCR) measures the number of times the signal waveform changes sign in the course of the current frame and is given by

$$ZCR_r = \frac{1}{2} \sum_{n=1}^N | \text{sign}(x_r(n)) - \text{sign}(x_{r-1}(n)) | \quad (2)$$

where,

$$\text{sign}(x) = \begin{cases} 1, & x \geq 0; \\ -1, & x < 0. \end{cases}$$

For several applications, the ZCR provides spectral information at a low cost. For narrowband signals (e.g. a sinusoid), the ZCR is directly related to the fundamental frequency. For more complex signals, the ZCR correlates well with the average frequency of the major energy concentration. For speech signals, the short-time ZCR takes on values that fluctuate rapidly between voiced and unvoiced segments due to their differing spectral energy concentrations. For music signals, on the other hand, the ZCR is more stable across extended time durations.

### Short-Time Energy

It is the mean squared value of the waveform values in the data frame and represents the temporal envelope of the signal. More than its actual magnitude, its variation over time can be a strong indicator of underlying signal content. It is computed as

$$E_r = \frac{1}{N} \sum_{n=1}^N |x_r(n)|^2 \quad (3)$$

### Band-Level Energy

It refers to the energy within a specified frequency region of the signal spectrum. It can be computed by the appropriately weighted summation of the power spectrum as given by

$$E_r = \frac{1}{N} \sum_{k=1}^{\frac{N}{2}} (X_r[k]W[k])^2 \quad (4)$$

$W[k]$  is a weighting function with non-zero values over only a finite range of bin indices “ $k$ ” corresponding to the frequency band of interest. Sudden transitions in the band-level energy indicate a change in the spectral energy distribution, or timbre, of the signal, and aid in audio segmentation. Generally log transformations of energy are used to improve the spread and represent (the perceptually more relevant) relative differences.

### Spectral Centroid

It is the center of gravity of the magnitude spectrum. It is a gross indicator of spectral shape. The spectral centroid frequency location is high when the high frequency content is greater.

$$C_r = \frac{\sum_{k=1}^{\frac{N}{2}} f[k] |X_r[k]|}{\sum_{k=1}^{\frac{N}{2}} |X_r[k]|} \quad (5)$$

Since moving the major energy concentration of a signal towards higher frequencies makes it sound brighter, the spectral centroid has a strong correlation to the subjective sensation of brightness of a sound [10].

### Spectral Roll-off

It is another common descriptor of gross spectral shape. The roll-off is given by

$$R_r = f[K] \quad (6)$$

where  $K$  is the largest bin that fulfills

$$\sum_{k=1}^K |X_r[k]| \leq 0.85 \sum_{k=1}^{\frac{N}{2}} |X_r[k]| \quad (7)$$

That is, the roll-off is the frequency below which 85% of accumulated spectral magnitude is concentrated. Like the centroid, it takes on higher values for right-skewed spectra.

### Spectral Flux

It is given by the frame-to-frame squared difference of the spectral magnitude vector summed across frequency as

$$F_r = \sum_{k=1}^{\frac{N}{2}} (|X_r[k]| - |X_{r-1}[k]|)^2 \quad (8)$$

It provides a measure of the local spectral rate of change. A high value of spectral flux indicates a sudden change in spectral magnitudes and therefore a possible segment boundary at the  $r^{th}$  frame.

### Fundamental Frequency ( $F_0$ )

It is computed by measuring the periodicity of the time-domain waveform. It may also be estimated from the signal spectrum as the frequency of the first harmonic or as the spacing between harmonics of the periodic signal. For real musical instruments and the human voice,  $F_0$  estimation is a non-trivial problem due to (i) period-to-period variations of the waveform, and (ii) the

fact that the fundamental frequency component may be weak relative to the other harmonics. The latter causes the detected period to be prone to doubling and halving errors. Time-domain periodicity can be estimated from the signal autocorrelation function (ACF) given by

$$R(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} (x_r[n]x_r[n + \tau]) \quad (9)$$

The ACF,  $R(\tau)$ , will exhibit local maxima at the pitch period and its multiples. The fundamental frequency of the signal is estimated as the inverse of the lag “ $\tau$ ” that corresponds to the maximum of  $R(\tau)$  within a predefined range. By favouring short lags over longer ones, fundamental frequency multiples are avoided. The normalized value of the lag at the estimated period represents the strength of the signal periodicity and is referred to as the harmonicity coefficient.

### Mel-Frequency Cepstral Coefficients (MFCC)

MFCC are perceptually motivated features that provide a compact representation of the short-time spectrum envelope. MFCC have long been applied in speech recognition and, much more recently, to music [11]. To compute the MFCC, the windowed audio data frame is transformed by a DFT. Next, a Mel-scale filterbank is applied in the frequency domain and the power within each sub-band is computed by squaring and summing the spectral magnitudes within bands. The Mel-frequency scale, a perceptual scale like the critical band scale, is linear below 1 kHz and logarithmic above this frequency. Finally the logarithm of the bandwise power values are taken and decorrelated by applying a DCT to obtain the cepstral coefficients. The log transformation serves to deconvolve multiplicative components of the spectrum such as the source and filter transfer function. The decorrelation results in most of the energy being concentrated in a few cepstral coefficients. For instance, in 16 kHz sampled speech, 13 low-order MFCCs are adequate to represent the spectral envelope across phonemes.

A related feature is the cepstral residual computed as the difference between the signal spectrum and the spectrum reconstructed from the prominent low-order cepstral coefficients. The cepstral residual thus provides a measure of the fit of the cepstrally smoothed spectrum to the spectrum.

### Feature Summaries

All the features described so far were short-time parameters computed at the frame rate from windowed segments of audio of duration no longer than 40 ms, the assumed interval of audio signal stationarity. An equally important cue to signal identity is the temporal pattern of changing signal properties observed across a sufficiently long interval. Local temporal changes may be described by

the time derivatives of the features known as delta-features. Texture windows, as indicated in Figure 3, enable descriptions of the long-term characteristics of the signal in terms of statistical measures of the time variation of each feature. Feature average and feature variance over the texture window serve as a coarse summary of the temporal variation. A more detailed description of a feature's variation with time is provided by the frequency spectrum of its temporal trajectory. The energy within a specific frequency band of this spectrum is termed a “modulation energy” of the feature [8]. For example, the short-time energy feature of speech signals shows high modulation energy in a band around 4 Hz due to the syllabic rate of normal speech utterances (that is, approximately 4 syllables are uttered per second leading to 4 local maxima in the short-time energy per second).

## 4.2 Perceptual Features

The human recognition of sound is based on the perceptual attributes of the sound. When a good source model is not available, perceptual features provide an alternative basis for segmentation and classification. The psychological sensations evoked by a sound can be broadly categorized as loudness, pitch and timbre. Loudness and pitch can be ordered on a magnitude scale of low to high. Timbre, on the other hand, is a more composite sensation with several dimensions that serves to distinguish different sounds of identical loudness and pitch. A computational model of the auditory system is used to obtain numerical representations of short-time perceptual parameters from the audio waveform segment. Loudness and pitch together with their temporal fluctuations are common perceptual features and are briefly reviewed here.

### Loudness

Loudness is a sensation of signal strength. As would be expected it is correlated with the sound intensity, but it is also dependent on the duration and the spectrum of the sound. In physiological terms, the perceived loudness is determined by the sum total of the auditory neural activity elicited by the sound. Loudness scales nonlinearly with sound intensity. Corresponding to this, loudness computation models obtain loudness by summing the contributions of critical band filters raised to a compressive power [12]. Salient aspects of loudness perception captured by loudness models are the nonlinear scaling of loudness with intensity, frequency dependence of loudness and the additivity of loudness across spectrally separated components.

### Pitch

Although pitch is a perceptual attribute, it is closely correlated with the physical attribute of fundamental frequency ( $F_0$ ). Subjective pitch changes

are related to the logarithm of  $F_0$  so that a constant pitch change in music refers to a constant ratio of fundamental frequencies. Most pitch detection algorithms (PDAs) extract  $F_0$  from the acoustic signal, i.e. they are based on measuring the periodicity of the signal via the repetition rate of specific temporal features, or by detecting the harmonic structure of its spectrum. Auditorily motivated PDAs use a cochlear filterbank to decompose the signal and then separately estimate the periodicity of each channel via the ACF [13]. Due to the higher channel bandwidths in the high frequency region, several higher harmonics get combined in the same channel and the periodicity detected then corresponds to that of the amplitude envelope beating at the fundamental frequency. The perceptual PDAs try to emulate the ear's robustness to interference-corrupted signals, as well as to slightly anharmonic signals, which still produce a strong sensation of pitch. A challenging problem for PDAs is the pitch detection of a voice when multiple sound sources are present as occurs in polyphonic music.

As in the case of the physical features, temporal trajectories of pitch and loudness over texture window durations can provide important cues to sound source homogeneity and recognition. Modulation energies of bandpass filtered audio signals, corresponding to the auditory gammatone filterbank, in the 20-40 Hz range are correlated with the perceived roughness of the sound while modulation energies in the 3-15 Hz range are indicative of speech syllabic rates [8].

## 5 Audio Classification Systems

We review a few prominent examples of audio classification systems. Speech and music dominate multimedia applications and form the major classes of interest. As mentioned earlier, the proper design of the feature set considering the intended audio categories is crucial to the classification task. Features are chosen based on the knowledge of the salient signal characteristics either in terms of production or perception. It is also possible to select features from a large set of possible features based on exhaustive comparative evaluations in classification experiments. Once the features are extracted, standard machine learning techniques to design the classifier. Widely used classifiers include statistical pattern recognition algorithms such as the  $k$  nearest neighbours, Gaussian classifier, Gaussian Mixture Model (GMM) classifiers and neural networks [14]. Much of the effort in designing a classifier is spent collecting and preparing the training data. The range of sounds in the training set should reflect the scope of the sound category. For example, car horn sounds would include a variety of car horns held continuously and also as short hits in quick succession. The model extraction algorithm adapts to the scope of the data and thus a narrower range of examples produces a more specialized classifier.

## 5.1 Speech-Music Discrimination

Speech-music discrimination is considered a particularly important task for intelligent multimedia information processing. Mixed speech/music audio streams, typical of entertainment audio, are partitioned into homogenous segments from which non-speech segments are separated. The separation would be useful for purposes such as automatic speech recognition and text alignment in soundtracks, or even simply to automatically search for specific content such as news reports among radio broadcast channels. Several studies have addressed the problem of robustly distinguishing speech from music based on features computed from the acoustic signals in a pattern recognition framework. Some of the efforts have applied well-known features from statistical speech recognition such as LSFs and MFCC based on the expectation that their potential for the accurate characterization of speech sounds would help distinguish speech from music [11], [15]. Taking the speech recognition approach further, Williams and Ellis [16] use a hybrid connectionist-HMM speech recogniser to obtain the posterior probabilities of 50 phone classes from a temporal window of 100 ms of feature vectors. Viewing the recogniser as a system of highly tuned detectors for speech-like signal events, we see that the phone posterior probabilities will behave differently for speech and music signals. Various features summarizing the posterior phone probability array are shown to be suitable for the speech-music discrimination task.

A knowledge-based approach to feature selection was adopted by Scheirer and Slaney [17], who evaluated a set of 13 features in various trained-classifier paradigms. The training data, with about 20 minutes of audio corresponding to each category, was designed to represent as broad a class of signals as possible. Thus the speech data consisted of several male and female speakers in various background noise and channel conditions, and the music data contained various styles (pop, jazz, classical, country, etc.) including vocal music. Scheirer and Slaney [17] evaluated several of the physical features, described in Sec. 4.1, together with the corresponding feature variances over a one-second texture window. Prominent among the features used were the spectral shape measures and the 4 Hz modulation energy. Also included were the cepstral residual energy and, a new feature, the pulse metric. Feature variances were found to be particularly important in distinguishing music from speech. Speech is marked by strongly contrasting acoustic properties arising from the voiced and unvoiced phone classes. In contrast to unvoiced segments and speech pauses, voiced frames are of high energy and have predominantly low frequency content. This leads to large variations in ZCR, as well as in spectral shape measures such as centroid and roll-off, as voiced and unvoiced regions alternate within speech segments. The cepstral residual energy too takes on relatively high values for voiced regions due to the presence of pitch pulses. Further the spectral flux varies between near-zero values during steady vowel regions to high values during phone transitions while that for music is more steady. Speech segments also have a number of quiet or low energy

frames which makes the short-time energy distribution across the segment more left-skewed for speech as compared to that for music. The pulse metric (or “rhythmicness”) feature is designed to detect music marked by strong beats (e.g. techno, rock). A strong beat leads to broadband rhythmic modulation in the signal as a whole. Rhythmicness is computed by observing the onsets in different frequency channels of the signal spectrum through band-pass filtered envelopes. There were no perceptual features in the evaluated feature set. The system performed well (with about 4% error rate), but not nearly as well as a human listener. Classifiers such as k-nearest neighbours and GMM were tested and performed similarly on the same set of features suggesting that the type of classifier and corresponding parameter settings was not crucial for the given topology of the feature space.

Later work [18] noted that music dominated by vocals posed a problem to conventional speech-music discrimination due to its strong speech-like characteristics. For instance, MFCC and ZCR show no significant differences between speech and singing. Dynamic features prove more useful. The 4 Hz modulation rate, being related to the syllabic rate of normal speaking, does well but is not sufficient by itself. The coefficient of harmonicity together with its 4 Hz modulation energy better captures the strong voiced-unvoiced temporal variations of speech and helps to distinguish it from singing. Zhang and Kuo [19] use the shape of the harmonic trajectories (“spectral peak tracks”) to distinguish singing from speech. Singing is marked by relatively long durations of continuous harmonic tracks with prominent ripples in the higher harmonics due to pitch modulations by the singer. In speech, harmonic tracks are steady or slowly sloping during the course of voiced segments, interrupted by unvoiced consonants and by silence. Speech utterances have language-specific basic intonation patterns or pitch movements for sentence clauses.

## 5.2 Audio Segmentation and Classification

Audiovisual data, such as movies or television broadcasts, are more easily navigated using the accompanying audio rather than by observing visual clips. Audio clips provide easily interpretable information on the nature of the associated scene such as for instance, explosions and shots during scenes of violence where the associated video itself may be fairly varied. Spoken dialogues can help to demarcate semantically similar material in the video while a continuous background music would help hold a group of seemingly disparate visual scenes together. Zhang and Kuo [19] proposed a method for the automatic segmentation and annotation of audiovisual data based on audio content analysis. The audio record is assumed to comprise of the following non-simultaneously occurring sound classes: silence, sounds with and without music background including the sub-categories of harmonic and inharmonic environmental sounds (e.g. touch tones, doorbell, footsteps, explosions). Abrupt changes in the short-time physical features of energy, zero-crossing rate and fundamental frequency are used to locate segment boundaries between the



distinct sound classes. The same short-time features, combined with their temporal trajectories over longer texture windows, are subsequently used to identify the class of each segment. To improve the speech-music distinction, spectral peaks detected in each frame are linked to obtain continuous spectral peak tracks. While both speech and music are characterized by continuous harmonic tracks, those of speech correspond to lower fundamental frequencies and are shorter in duration due to the interruptions from the occurrence of unvoiced phones and silences.

Wold et al [20] in a pioneering work addressed the task of finding similar sounds in a database with a large variety of sounds coarsely categorized as musical instruments, machines, animals, speech and sounds in nature. The individual sounds ranged in duration from 1 to 15 seconds. Temporal trajectories of short-time perceptual features such as loudness, pitch and brightness were examined for sudden transitions to detect class boundaries and achieve the temporal segmentation of the audio into distinct classes. The classification itself was based on the salient perceptual features of each class. For instance, tones from the same instrument share the same quality of sound, or timbre. Therefore the similarity of such sounds must be judged by descriptors of temporal and spectral envelope while ignoring pitch, duration and loudness level. The overall system uses the short-time features of pitch, amplitude, brightness and bandwidth, and their statistics (mean, variance and autocorrelation coefficients) over the whole duration of the sound.

### 5.3 Music Information Retrieval

Automatically extracting musical information from audio data is important to the task of structuring and organizing the huge amounts of music available in digital form on the Web. Currently, music classification and searching depends entirely upon textual meta-data (title of the piece, composer, players, instruments, etc.). Developing features that can be extracted automatically from recorded audio for describing musical content would be very useful for music classification and subsequent retrieval based on various user-specified similarity criteria. The chief attributes of a piece of music are its timbral texture, pitch content and rhythmic content. The timbral texture is related to the instrumentation of the music and can be a basis for similarity between music drawn from the same period, culture or geographical region. Estimating the timbral texture would also help to detect a specified solo instrument in the audio record. The pitch and rhythm aspects are linked to the symbolic transcription or the “score” of the music independent of the instruments playing. The pitch and duration relationships between successive notes make up the melody of the music. The rhythm describes the timing relation between musical events within the piece including the patterns formed by the accent and duration of the notes. The detection of note onsets based on low-level audio features is a crucial component of rhythm detection. Automatic music transcription, or the conversion of the sound signal to readable musical nota-

tion, is the starting point for retrieval of music based on melodic similarity. As a tool, it is valuable also to music teaching and musicological research. A combination of low-level features for each of higher-level musical attributes viz. timbre, rhythmic structure and pitch content are used in identification of musical style, or genre [9], [10]. We next discuss the various features used for describing musical attributes for specific applications in music retrieval.

Timbral texture of music is described by features similar to those used in speech and speaker recognition. Musical instruments are modeled as resonators with periodic excitation. For example, in a wind instrument, a vibrated reed delivers puffs of air to a cylindrical bore resonator. The fundamental frequency of the excitation determines the perceived pitch of the note while the spectral resonances, shaping the harmonic spectrum envelope, are characteristic of the instrument type and shape. MFCC have been very successful in characterizing vocal tract resonances for vowel recognition, and this has prompted their use in instrument identification. Means and variances of the first few cepstral coefficients (excluding the DC coefficient) are utilized to capture the gross shape of the spectrum [9]. Other useful spectral envelope descriptors are the means and variances over a texture window of the spectral centroid, roll-off, flux and zero-crossing rate. The log of the attack time (duration between the start of the note and the time at which it reaches its maximum value) together with the energy-weighted temporal centroid are important temporal descriptors of instrument timbre especially for percussive instruments [21]. Given the very large number of different instruments, a hierarchical approach to classification is sometimes taken based on instrument taxonomy.

A different, and possibly more significant, basis for musical similarity is the melody. Melodic similarity is an important aspect in music copyright and detection of plagiarism. It is observed that in the human recall of music previously heard, the melody or tune is by far the most well preserved aspect [22]. This suggests that a natural way for a user to query a music database is to perform the query by singing or humming a fragment of the melody. “Query-by-humming” systems allow just this. Users often prefer humming the tune in a neutral syllable (such as “la” or “ta”) to singing the lyrics. Retrieving music based on a hummed melody fragment then reduces to matching the melodic contour extracted from the query with pre-stored contours corresponding to the database. The melodic contour is a mid-level data representation derived from low-level audio features. The mid-level representation selected defines a trade-off between retrieval accuracy and robustness to user errors. Typically, the melody is represented as a sequence of discrete-valued note pitches and durations. A critical component of query signal processing then is audio segmentation into distinct notes, generally considered a challenging problem [23]. Transitions in short-term energy or in band-level energies derived from either auditory or acoustic-phonetic motivated frequency bands have been investigated to detect vocal note onsets [23], [24]. A pitch label is assigned to a detected note based on suitable averaging of frame-level pitch estimates across

the note duration. Since people remember the melodic contour (or the shape of the temporal trajectory of pitch) rather than exact pitches, the note pitches and durations are converted to relative pitch and duration intervals. In the absence of an adequate cognitive model for melodic similarity, modifications of the basic string edit distance measure are applied to template matching in the database search [26].

The instrument identification and melody recognition methods discussed so far implicitly assumed that the music is monophonic. In polyphonic music, where sound events corresponding to different notes from one or more instruments overlap in time, pitch and note onset detection become considerably more difficult. Based on the observation that humans can decompose complex sound mixtures to perceive individual characteristics such as pitch, auditory models are being actively researched for polyphonic music transcription [13], [25].

## 6 Summary

The world's ever growing archives of multimedia data pose huge challenges for digital content management and retrieval. While we now have the ability to search text quite effectively, other multimedia data such as audio and video remain opaque to search engines except through a narrow window provided by the possibly attached textual metadata. Segmentation and classification of raw audio based on audio content analysis constitutes an important component of audiovisual data handling. Speech and music constitute the commercially most significant components of audio multimedia. While speech and speaker recognition are relatively mature fields, music information retrieval is a new and growing research area with applications in music searching based on various user-specified criteria such as style, instrumentation and melodic similarity. In this chapter we have focused on signal processing methods for the segmentation and classification of audio signals. Assigning class labels to sounds or audio segments is aided by an understanding of either source or signal characteristics. The general characteristics of audio signals have been discussed, followed by a review of time-frequency representations that help in the visualization of the acoustic spectro-temporal properties. Since the important attributes of an audio signal are its salient features as perceived by the human ear, auditory models can play a substantial role in the effective representation of audio. Two chief components of a classification system are the feature extraction module and the classifier itself. The goal is to find features that are invariant to irrelevant transformations and have good discriminative power across the classes. Feature extraction, an important signal processing task, is the process of computing a numerical representation from the acoustical signal that can be used to characterize the audio segment. Important audio features, representative of the rich library developed by research in audio classification over the years, have been reviewed. The acoustic and

perceptual correlates of the individual features have been discussed, providing a foundation for feature selection in specific applications.

Local transitions in feature values provide the basis for audio segmentation. Sound tracks could, for instance, be comprised of alternating sequences of silence, spoken dialog and music needing to be separated into homogenous segments before classification. Classification algorithms typically use labeled training examples to partition the feature space into regions so that feature vectors falling in the same region come from the same class. Statistical pattern classifiers and neural networks are widely used. Since the methods are well documented in the general machine learning literature, they have not been discussed here. A few important audio classification tasks have been presented together with a discussion of some widely cited proposed solutions. The problems presented include speech-music discrimination, audiovisual scene segmentation and classification based on audio features, and music retrieval applications such as genre or style identification, instrument identification and query by humming. Active research continues on all these problems for various database compilations within the common framework of feature extraction followed by pattern classification. Significant progress has been made and improvements in classification accuracy continue to be reported from minor modifications to the basic low-level features or in the classification method employed. However it is believed that such an approach has reached a “glass ceiling”, and that really large gains in the performance of practical audio search and retrieval systems can come only with a deeper understanding of human sound perception and cognition [27]. Achieving the separation of the components of complex sound mixtures, and finding objective distance measures that predict subjective judgments on the similarity of sounds are specific tasks that would benefit greatly from better auditory system modelling.

## References

1. Oppenheim A V and Lim J S, The Importance of Phase in Signals. Proc of the IEEE 69(5):529-550
2. Moore BCJ (2003) An Introduction to the Psychology of Hearing. Academic Press, San Diego
3. Patterson R D (2000) Auditory Images: How Complex Sounds Are Represented in the Auditory System. J Acoust Soc Japan (E) 21(4)
4. Lyon R F, Dyer L (1986) Experiments with a Computational Model of the Cochlea. Proc of the Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)
5. Martinez J M (2002) Standards - MPEG-7 overview of MPEG-7 description tools, part 2. IEEE Multimedia 9(3):83-93
6. Xiong Z, Radhakrishnan R, Divakaran A, Huang T (2003) Comparing MFCC and MPEG-7 Audio Features for Feature Extraction, Maximum Likelihood HMM and Entropic Prior HMM for Sports Audio Classification. Proc of the Intl Conf on Multimedia and Expo (ICME)

7. Wang L, Brown G (2006) Computational Auditory Scene Analysis: Principles, Algorithms and Applications. Wiley-IEEE Press, New York
8. McKinney M F, Breebaart J (2003) Features for Audio and Music Classification. Proc of the Intl Symp on Music Information Retrieval (ISMIR)
9. Tzanetakis G, Cook P (2002) Musical Genre Classification of Audio Signals. IEEE Trans on Speech and Audio Processing 10(5):293-302
10. Burred J J, Lerch A (2004) Hierarchical Automatic Audio Signal Classification. J Audio Engineering Society 52(7/8):724-739
11. Logan B (2000) Mel frequency cepstral coefficients for music modeling. Proc of the Intl Symp on Music Information Retrieval (ISMIR)
12. Zwicker E, Scharf B (1965) A Model of Loudness Summation. Psychological Review 72:3-26
13. Klapuri A P (2005) A Perceptually Motivated Multiple-F0 Estimation Method for Polyphonic Music Signals. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPA)
14. Duda R, Hart P, Stork D (2000) Pattern Classification. Wiley, New York
15. El-Maleh K, Klein M, Petrucci G, Kabal P (2000) Speech/music Discrimination for Multimedia Applications. Proc of the Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)
16. Williams G, Ellis D (1999) Speech/music Discrimination based on Posterior Probability Features. Proc of Eurospeech
17. Scheirer E, Slaney M (1997) Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. Proc of the Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)
18. Chou W, Gu L (2001) Robust Singing Detection in Speech/Music Discriminator Design. Proc of the Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)
19. Zhang T, Kuo C C J (2001) Audio Content Analysis for Online AudioVisual Data Segmentation and Classification. IEEE Trans on Speech and Audio Processing 9(4):441-457
20. Wold E, Blum T, Keisler D, Wheaton J (1996) Content-based Classification, Search and Retrieval of Audio. IEEE Multimedia 3(3):27-36
21. Peeters G, McAdams S, Herrera P (2000) Instrument Sound Description in the Context of MPEG-7. Proc of the Intl Computer Music Conference (ICMC)
22. Dowling W J (1978) Scale and Contour: Two Components of a Theory of Memory for Melodies. Psychological Review 85:342-389
23. Pradeep P, Joshi M, Hariharan S, Dutta-Roy S, Rao P (2007) Sung Note Segmentation for a Query-By-Humming System. Proc of the Intl Workshop on Artificial Intelligence and Music (Music-AI) in IJCAI
24. Klapuri A P (1999) Sound Onset Detection by Applying Psychoacoustic Knowledge. Proc of the Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)
25. de Cheveigne A, Kawahara H (1999) Multiple period estimation and pitch perception model. Speech Communication 27:175-185
26. Uitdenbogerd A, Zobel J (1999) Melodic Matching Techniques for Large Music Databases. Proc of the 7th ACM Intl Conference on Multimedia (part 1)
27. Aucouturier J J, Pachet F (2004) Improving Timbre Similarity: How High is the Sky. J Negative Results in Speech and Audio Sciences 1(1)