SPRINGER BRIEFS IN ELECTRICAL AND COMPUTER ENGINEERING · SPEECH TECHNOLOGY

Mohamed Hesham Farouk

Application of Wavelets in Speech Processing

Springer

SpringerBriefs in Electrical and Computer Engineering

Speech Technology

Series editor

Amy Neustein, Fort Lee, USA

For further volumes:
http://www.springer.com/series/10043

Mohamed Hesham Farouk

Application of Wavelets in Speech Processing

Springer

Mohamed Hesham Farouk
Engineering Mathematics and Physics
Department
Cairo University
Cairo
Egypt

ISSN 2191-737X ISSN 2191-7388 (electronic)
ISBN 978-3-319-02731-9 ISBN 978-3-319-02732-6 (eBook)
DOI 10.1007/978-3-319-02732-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013955564

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To the soul of my mother

Contents

Chapter 1
Introduction

Abstract As the wavelets gain wide applications in different fields, especially within the signal processing realm, this chapter will provide a survey on widespread employing of wavelets analysis in different applications of speech processing. Many speech processing algorithms and techniques still lack some sort of robustness which can be improved through the use of wavelet tools. Researchers and practitioners in speech technology will find valuable information on the use of wavelets to strengthen both development and research in different applications of speech processing.

Keywords Wavelet transform · Speech processing · Applications in speech technology · Wavelet spectral analysis · Wavelet-basis functions

In this monograph, we discuss many proposed algorithms which employ wavelet transform (WT) for different applications in speech technology. A survey was conducted through recent works which used WT in speech processing realms. This survey covers both the use of wavelets in enhancing previously proposed algorithms and new algorithms based, principally, on wavelet analysis. In general, wavelet analysis can serve through many ways in speech processing since it can provide new enhanced spectral analysis approach, basis-expansion for signals, identification features, and can serve well for noise cancellation.

1.1 History and Definition of Speech Processing

The first trials for speech processing through machines are dated to the ancient Egyptians who built statutes producing sound. There are other documents belonging to the eighteenth century, which attempt at building speaking machines [1].

In human speech processing system, several transformations may be included, such as thought-to-articulation, articulators movement to acoustical signal, propagation of the speech signal, electronic transmission/storage, loudspeaker to the listener's ears, acoustic to electrical in the inner ear, and interpretation by the listener's brain. These transformations are modeled through many mathematical

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_1,

algorithms. In most speech processing algorithms, a feature space is built based on a transformation kernel to a space of lower dimension, which allows a post-processing stage, readily, resulting in more useful information.

Accordingly, speech processing is discussing the methods and algorithms used in analyzing and manipulating speech signals. Since the signals are usually processed in a digital representation, hence speech processing can be regarded as a special case of digital signal processing, applied to speech signal. The main topics of speech processing are recognition, synthesis, enhancement, and coding. The processing for speech recognition is concentrated on extracting the best features that can achieve highest recognition rate using a certain classifier. For speech coding and synthesis, the coding parameters from speech signals should result in a closest matching between the original and reconstructed signals. While for speech enhancement, efforts are directed toward discovering analysis components that may comprise sources of signal degradation.

1.2 Applications of Speech processing

Applications of speech processing techniques may include compression and enhancement of human speech, clinical diagnosis of speech disorders, man–machine interfacing through voice, security systems for speech communications, machine translation of speech, reading machines, and understanding based on voice communication. In these applications, the speech signal is, customarily, transformed from time domain to another domain in which efficient features can be extracted to express a functional character of such signal. As an example, spectral features can be exploited in identifying meaning or type of a speech signal in the field of speech and speaker recognitions. For other applications, the features can be used for reconstructing the signal again after analysis and enhancement.

Many real-world applications are now available based on research results in speech processing. It may range from the use of speech-based dictation machines to speech-based command and control supporting interactions with machines. Speech recognition can serve as an alternative interface to the traditional one or be a complement modality speeding up analysis process and increasing its fidelity and convenience. Many applications for critical society services are continuously being improved and are made easily accessible through a speech-based interface. Cellular phones and multimedia systems also employ speech-coding algorithms. Even diagnosis of speech disorder can benefit from results obtained by research in speech processing. Speech processing also find applications in security fields [2].

1.3 Recent Progress in Speech Processing

Most applications of speech processing have emerged many years ago. However, recent years have seen the widespread deployment of smart phones and other portable

devices with the ability to make good quality recordings of speech and even video. Such recordings can be processed locally or transmitted for processing on remote systems having more computational power and storage. More computational power and storage increase rapidly day by day as computational technology advances.

The current state of speech processing systems is still far from human performance. A major problem for most speech-based applications is robustness, which refers to the fact that they may be insufficiently general. As an example, a truly robust ASR system should be independent of any speaker, in reasonable environments. Environmental noise, from natural sources or machines, as well as communication channel distortions, all tend to degrade the system's performance, often severely. Human listeners, by contrast, can often adapt rapidly to these difficulties, which suggests that there remain significant enhancement needed. However, much of what we know about human speech production and perception need to be integrated into research efforts in the near future.

As a result, the main objective of research in speech processing is directed toward finding techniques for robust speech processing. This concern has been motivated by the increase in need for low complexity and efficient speech feature extraction methods, the need for enhancing the naturalness, acceptability, and intelligibility of the reconstructed speech signal corrupted by environmental noise, and the need for reducing noise for robust speech recognition systems to achieve high recognition rate in harsh environments [3]. New algorithms are continuously developed for enhancing the performance of speech processing for different applications. Most improvements are founded on the ever growing research infrastructure in speech area and inspiration from related biological systems. The technology of powerful computation and communication systems admits more sophisticated and efficient algorithms to be employed for reliable and robust applications of speech processing. Besides such systems, larger speech corpora are available and, in general, the infrastructure of research in speech area is growing to be more helpful.

Eventually, different merits of WT can serve as efficient features in, approximately, most research concerns, especially with newer versions of speech corpora emerging continuously.

References

1. J. Benesty, M. Sondhi, Y. Huang, in *Springer Handbook of Speech Processing*. (Springer, New York, 2007)
2. R. Hu, S. Zhu, J. Feng, A. Sears, in *Use of Speech Technology in Real Life Environment, Lecture Notes in Computer Science*, vol. 6768. (Springer, New York, 2011), pp. 62–71
3. V.T. Pham, Wavelet analysis for robust speech processing and applications, Ph.D. dissertation, 2007, http://theses.eurasip.org/media/theses/documents/pham-van-tuan-wavelet-analysis-for-robust-speech-processing-and-applications.pdf

Chapter 2
Speech Production and Perception

Abstract The main objective of research in speech processing is directed toward finding techniques for extracting features, which robustly model a speech signal. Some of these features can be characterized by relatively simple models, while others may require more realistic models in both cases of speech production and perception.

Keywords Speech production modeling · Spectral analysis of speech · Wavelet transform · Speech perception

Speech sounds are produced through the movement of organs constituting the vocal tract (glottis, velum, tongue, lips) acting on the air in the respiratory passages (trachea, larynx, pharynx, mouth, nose). The vocal organs generate a local disturbance on the air at several positions in the vocal tract creating the sources for speech production. The acoustic waves generated by such sources are then modulated during the propagation through the vocal tract with a specific shape. Accordingly, the structure of speech sounds is generated by the combined effect of sound sources and vocal tract characteristics. The source-filtering model of speech production assums that the spectrum of source excitation at glottis is then shaped according to filtering properties of vocal tract. Such filtering properties are changing continuously with time. The continuous changes in the shape of vocal tract and excitations through glottis make the produced sounds at lips nonstationary. Wavelet analysis is one of the best methods for extracting spectral features from nonstationary signals, since it employs multiresolution measures both in time and frequency.

2.1 Speech Production Process

The speech production process takes place inside the vocal tract extending from the glottis to the lips. The process is energized from air-filled lungs. The vocal tract is a chamber of extremely complicated geometrical shape whose dimensions and configuration may vary continuously in time and whose walls are composed of tissues having widely ranging properties. It begins at the glottis and ends at the lips.

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_2,

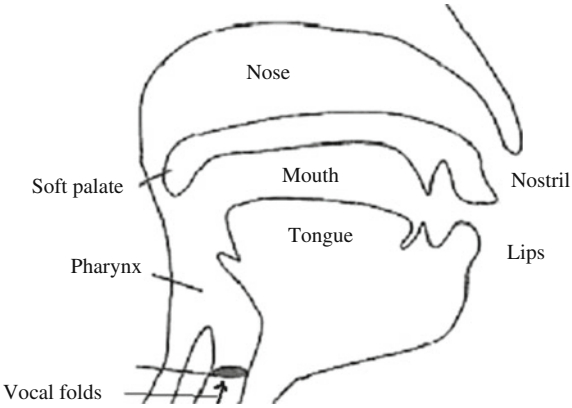

Fig. 2.1 Anatomical structure of vocal tract

Figure 2.1 shows the vocal tract anatomical structure. The glottis is a slit-like orifice between the vocal cords (at the top of the trachea). The cartilages around the cords support them and facilitate adjustment of their tension. The flexible structure of the vocal cords makes them oscillate easily. These oscillations are responsible for periodic excitation of vowels. The nasal tract constitutes an ancillary path for sound transmission. It begins at the velum and terminates at the nostrils.

2.2 Classification of Speech Sounds

Speech sounds are classified according to the type and place of excitation. Voiced sounds and vowels are characterized by a periodic excitation at the glottis. For voiced sounds and vowels, the expelled air from lungs causes the vocal cords to vibrate as a relaxation oscillator, and the air stream is modulated into discrete puffs. This oscillation starts when the subglottal pressure is increased sufficiently to force the initially abducted cords apart with lateral acceleration. As the air flow builds up in the orifice, the local pressure is reduced and a force acts to return the cords to a proximate position. Consequently, the pressure approaches the subglottal value as the flow decreases with the decrease in the orifice (glottal area). The relaxation cycle is then repeated. The mass and compliance of the cords, and the subglottal pressure determine the oscillation frequency (pitch) [1].

Unvoiced sounds are generated by passing the air stream through a constriction in the tract. The pressure perturbations due to these excitation mechanisms provide an acoustic wave which propagates along the vocal tract toward the lips.

If the nasal tract is coupled to the vocal cavity through the velum, the radiated sound is the resultant of the radiation at both the lips and nostrils and it is called nasalized sounds (as in/m/and/n/). The distinctive sounds of any language (phonemes) are uniquely determined by describing the excitation source and the vocal tract configuration.

The variation of the cross-sectional area (c.s.a.) along the vocal tract is called the *area function* according to the articulators positions. The area function of the vowels is determined primarily by the position of the tongue, but the positions of the jaw, lips, and, to a small extent, the velum also influence the resulting sound. The area function with the excitation type can uniquely define the produced sound.

2.3 Modeling of Speech Production

As discussed in the previous section, speech wave production can be divided into three stages: sound source generation, articulation by vocal tract, and radiation from the lips and/or nostrils.

Specifically, sound sources are either voiced or unvoiced. A voiced source can be modeled, in the simplest case, by a generator of periodic pulses or asymmetrical triangular waves which are repeated at every fundamental period (pitch). The peak value of the source wave corresponds to the loudness of the voice. An unvoiced sound source, on the other hand, can be modeled by a white noise generator, the mean energy of which corresponds to the loudness of voice [2].

For many speech applications such as coding, synthesis, and recognition, good performance can be achieved with a speech model that reflects broad characteristics of timing and articulatory patterns as well as varying frequency properties [3, 4]. In such a model, a scheme is designed to perform some spectral shaping on certain excitation wave so that it matches the natural spectrum, i.e., the vocal tract tube looks as a spectral shaper of the excitation. This approach is called "terminal analog" since its output is analogous to the natural process on the terminals only. The main interest in such approach is centered on resonance frequencies (formants or system poles) and their bandwidths. The two widely used methods of this approach are formant model [5] and linear prediction (LP) model [6]. These models provide simpler implementation schemes, both hardware and software. Many commercial products now adopt such models in their operation [7]. The adoption of terminal analog models affords sufficient intelligibility for many applications along with fast response due to their simplicity and amenability to implementation through many available media. Apparently, the extracted features using such models can be considered as different forms of resonances or spectral content of a speech signal. The wavelets are considered one of the efficient methods for representing the spectrum of speech signals [8].

2.4 Speech Perception Modeling

The ear is the main organ in the process of speech perception. It consists of an outer part, middle part, and inner part. Figure 2.2 shows the structure of human auditory system. The outer ear main function is to catch sound waves; this is done

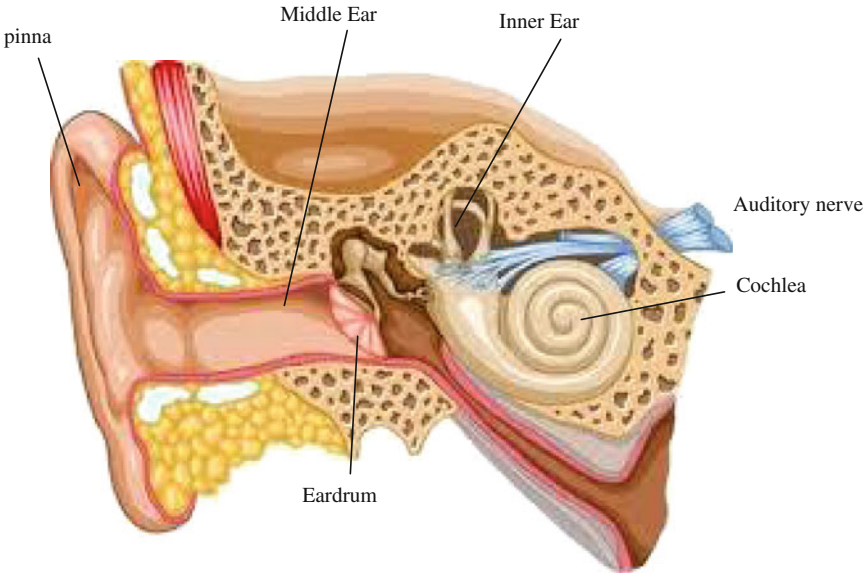

Fig. 2.2 Human auditory system

by the pinna. The pinna is pointed forward and has a number of curves to be able to catch the sound and determine its direction. After the sound reaches the pinna it is guided to the middle ear using the external auditory canal until it reaches the eardrum. The main function of the middle ear is to magnify the sound pressure and this is because the inner ear transfers sound through fluid and not air as in the middle and outer ear. Then the inner ear starts with the cochlea; the most important organ in human ear. The cochlea performs the spectral analysis of speech signal. The input stimulus is split into several frequency bands which are called critical bands [9]. The ear averages the energies of the frequencies within each critical band and thus forms a compressed representation of the original stimulus.

Studies have shown that human perception of the frequency content of sounds, either for pure tones or for speech signals, does not follow a linear scale. Majority of the speech and speaker recognition systems have used the feature vectors derived from a filter bank that has been designed according to the model of the auditory system. There are a number of forms used for these filters, but all of them are based on a frequency scale that is approximately linear below 1 kHz and approximately logarithmic above this point. Wavelet multiresolution analysis can provide accurate localization in both time and frequency domains which can emulate the human auditory system operation [8].

2.5 Intelligibility and Speech Quality Measures

The terms intelligibility and quality of speech are used interchangeably. The degradation of speech quality is mainly a result of background noise either through a communication channel or environment [3]. The evaluation of speech quality is highly important in many speech applications. Subjective listening or conversation tests are the most reliable measure of speech quality, however, these tests are often fairly expensive, time-consuming, labor intensive, and difficult to reproduce. However, for some applications like the assessment of alternative coding or enhancement algorithms, an objective measure is more economic to give the designer an immediate and reliable estimate of the anticipated perceptual quality of a particular algorithm. Traditional objective quality measures which rely on waveform matching like signal-to-noise ratio (SNR) or its variants like Segmental SNR (SSNR) are examples of straightforward measures. Perceptual quality measures are better candidates for fast assessment with more accurate results. The motivation for this perception-based approach is to create estimators which resemble that of human hearing system as described by the psychoacoustic models. In a psychoacoustic model of human hearing, the whole spectrum bandwidth of speech signal is divided into critical bands of hearing. The wavelet transform can be used for quality evaluation of speech in the context of critical band decomposition and auditory masking [10–12]. Moreover, wavelet analysis can reduce the computational effort associated with the mapping of speech signals into an auditory scale [10].

References

1. J. Flanagan, *Speech Analysis Synthesis and Perception* (Springer, New York, 1972)
2. M. Hesham, Vocal Tract Modeling, Ph. D. Dissertation, Faculty of Engineering, Cairo University, 1994
3. J. Deller, J. Proakis, J. Hansen, *Discrete-Time Processing of Speech Signals* (IEEE PRESS, New York, 2000)
4. M. Hesham, M.H. Kamel, A unified mathematical analysis for acoustic wave propagation inside the vocal tract. J. Eng. Appl. Sci. **48**(6), 1099–1114 (2001)
5. D. Klatt, Review of text-to-speech conversion for English. J. Acoust. Soc. Am. **82**(3), 737–793 (1987)
6. J.D. Markel, A.H.J. Gray, in *Linear Prediction of Speech (Chap. 4)*. (Springer, New York, 1976)
7. D. O'Shaughnessy, Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recogn. **41**(10), 2965–2979 (2008)
8. S. Ayat, A new method for threshold selection in speech enhancement by wavelet thresholding, in *International Conference on Computer Communication and Management (ICCCM 2011)*, Sydney, Australia, May 2011
9. J. Benesty, M. Sondhi, Y. Huang, in *Springer Handbook of Speech Processing*. (Springer, New York, 2007)

10. M. Hesham, A predefined wavelet packet for speech quality assessment. J. Eng. Appl. Sci. **53**(5), 637–652 (2006)
11. A. Karmakar, A. Kumar, R.K. Patney, A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality. IEEE Trans. Audio Speech Lang. Process. **14**(6), 1912–1923 (2006)
12. W. Dobson, J. Yang, K. Smart, F. Guo, High quality low complexity scalable wavelet audio coding, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Apr 1997, pp. 327–330

Chapter 3
Wavelets, Wavelet Filters, and Wavelet Transforms

Abstract Spectral characteristics of speech are known to be particularly useful in describing a speech signal such that it can be efficiently reconstructed after coding or identified for recognition. Wavelets are considered one of such efficient methods for representing the spectrum of speech signals. Wavelets are used to model both production and perception processes of speech. Wavelet-based features prove a success in a widespread area of practical applications in the speech processing realm.

Keywords Wavelet transform · Wavelet family · Wavelet filter · Multiresolution decomposition · Time–frequency analysis

Multiresolution analysis based on the wavelet theory permits the introduction of concepts of signal filtering with different bandwidths or frequency resolutions. The wavelet transform (WT) provides a framework to decompose a signal into a number of new signals, each one of them with different degrees of resolution. While the Fourier transform (FT) gives an idea of the frequency content in a signal, the wavelet representation is an intermediate representation between the Fourier and the time representation, and can provide good localization in both frequency and time domains. Fast variation in both domains can be detected by inspecting the coefficients of WT. Because of the difficult nature of speech signals and their fast variation with time, WT is used. In this part, we will review the properties of different approaches for obtaining WT.

3.1 Short-Time Fourier Transform (STFT)

In general, any mathematical transform for a signal or a function with time $s(t)$ takes the form

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_3,

$$S(\alpha) = \int_{-\infty}^{\infty} s(t)\ K(\alpha, t)\, dt \tag{3.1}$$

where $S(\alpha)$ is the transform of $s(t)$ with respect to the kernel $K(\alpha, t)$, and α is the transform variable. In Fourier transform, the kernel is $K(\omega, t) = e^{-j\omega t}$ where $\omega = 2\pi f$ is the angular frequency and f is the frequency.

FT is the main tool for spectral analysis of different signals. The short-time Fourier transform (STFT) cuts out a signal in short-time intervals (frames) and performs the Fourier Transform in order to capture time-dependent fluctuation of frequency component of a nonstationary signal. The STFT can be expressed as

$$S(\omega, \beta) = \int_{-\infty}^{\infty} s(t) \quad w(t - \beta)\, e^{-j\omega t}\, dt \tag{3.2}$$

where $s(t)$ is a signal, $S(\omega)$ is its STFT, and $w(t - \beta)$ is a window function centered around β in time. The window function is then shifted in time, and the Fourier transform of the product is computed again. So, for a fixed shift β of the window $w(t)$, the window captures the features of the signal $s(t)$ around different locations defined by β. The window helps to localize the time domain data within a limited period of time before obtaining the frequency domain information. The signal has been assumed quasistationary during the period of $w(t)$. The STFT can be viewed as a convolution of the signal $s(t)$ with a filter having an impulse response of the form $h(t) = w(t - \beta)\, e^{-j\omega t}$. The STFT can be also interpreted as a bank of narrow, slightly overlapping band-pass filters with additional phase information for each one. Alternatively, it can be seen as a special case of a family of transforms that use basis functions.

For STFT, in order to improve the accuracy with respect to time-dependent variation, it is necessary to shorten the frame period. The frequency resolution becomes worse with decreasing frame length. In other words, the requirements in the time localization and frequency resolution are conflicting.

The major drawback of the STFT is that it uses a fixed window width. Alternatively, the WT provides a better time–frequency representation of the signal than any other existing transforms. The WT solves the above problem to a certain extent. In contrast to STFT, which uses a single analysis window, the WT uses short windows at high frequencies and long windows at low frequencies.

3.2 Multiresolution Analysis and Wavelet Transform

In FT, a fixed window is used uniformly for a spread of frequencies; on the contrary, WT uses short windows at high frequencies and long windows at low frequencies. In this way, the characteristics of nonstationary speech signals can be

more closely examined. Accordingly, WT coefficients are localized in both time and frequency domains. This localization is constrained with Heisenberg's uncertainty principle which affirms that no transform can provide high resolution in both time and frequency domains at the same time. The useful locality property is exploited in this context. Because the wavelet basis functions are generated by scaling from a mother wavelet, they are well localized in time and scale domains. This behavior of wavelet decomposition is suitable for processing of speech signals which require high-frequency resolution to analyze low-frequency components (voiced sounds, formant frequencies), and high temporal resolution to analyze high-frequency components (mostly unvoiced sounds).

As a mathematical transform, the WT takes a kernel based on a function $\psi(t)$ as follows:

$$S(a,b) = \int_{-\infty}^{\infty} s(t) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt \tag{3.3}$$

where a is a scale parameter, b is another parameter for translation.

As in Eq. (3.3), the wavelet analysis is done similar to the STFT analysis except that the kernel function is not sinusoidal. The wavelet function $\psi(t)$ is a member in a wavelet family. A wavelet family is generated from what is called mother wavelet. All wavelets of a family share the same properties and their collection constitutes a complete basis.

The semi-discrete WT of the function $s(t)$, $s \in L^2(\Re)$ is defined as follows [1]:

Analysis equation

$$S(j,k) = \int_{-\infty}^{\infty} s(t) 2^{-j/2} \psi\left(2^{-j}t - kT_s\right) dt. \tag{3.4}$$

Synthesis or inverse-transform equation

$$s(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} S(j,k) 2^{-j/2} \psi^*\left(2^{-j}t - kT_s\right). \tag{3.5}$$

where j and k are indices indicating scale and location of a particular wavelet and without loss of generality, $T_s = 1$ can be considered in a discrete case.

The wavelet theory would immediately allow us to obtain line-frequency analysis and synthesis with the possibility to capture both long-lasting (low-frequency) components and to localize short irregularities, spikes, and other singularities with high-frequency content. The former objectives can be approached by wavelets at low scales, while the latter are successfully performed by wavelets at high scales and appropriate locations. Wavelet localization follows Heisenberg uncertainty principle in both the time and frequency domains that for any given wavelet $\Delta t \Delta f \geq 1/2\pi$.

As in Eq. (3.3), the pure wavelet expansion requests an infinite number of scales or resolutions k to represent the signal $s(t)$ completely. This is impractical. If the expansion is known only for certain scales $k < M$, we need a complement component to present information of expansion for $k > M$. This is done by introducing a scaling function $\varphi(t)$ such that [2],

$$\varphi_{j,k}(t) = 2^{-j/2}\varphi\left(2^{-j}t - k\right) \tag{3.6}$$

where the set $\varphi_{j,k}(t)$ is an orthonormal basis for subsapace of $L^2(\Re)$. With the introduced component, the signal $s(t)$ can be represented as a limit of successive approximations corresponding to different resolutions. This formulation is called a multiresolution analysis (MRA).

Consequently, the signal $s(t)$ can be set as the sum of an approximation plus M details at the Mth decomposed resolution or level. Equation (3.5) can be rewritten after including approximations as follows:

$$s(t) = \sum_k a_{M,k}\varphi_{M,k}(t) + \sum_{j=1}^{M}\sum_k d_{j,k}\psi_{j,k}(t) \tag{3.7}$$

where M represents the number of scales. $a_{M,k}$ are the approximation or scaling coefficients, and $d_{j,k}$ are the details or wavelet coefficients.

3.3 Wavelets and Bank of Filters

The WT can be viewed as a convolution of the signal and a wavelet function. In discrete-time domain, the set of discrete-time scaling and wavelet functions can be constructed from filter banks. The signal can be split into frequency bands through a bank of filters. As an example, a two-channel filter bank is shown in Fig. 3.1. The filters are therefore a low-pass $L(z)$ and a high-pass $H(z)$ filter.

The original signal can be reconstructed using such bank of filters. In the synthesis phase, the signals are upsampled and passed through the synthesis filters. The outputs of the filters in the synthesis bank are summed to get the reconstructed signal. The outputs of the analysis bank are called subbands and this technique is also called subband coding [3].

3.4 Wavelet Families

A wavelet function must be oscillatory in some way to capture a frequency band from the analyzed signal. The wavelet function comprises both the analyzing function and a short-time window. A wavelet family is generated from a mother function ψ_{jk} which can be defined as follows:

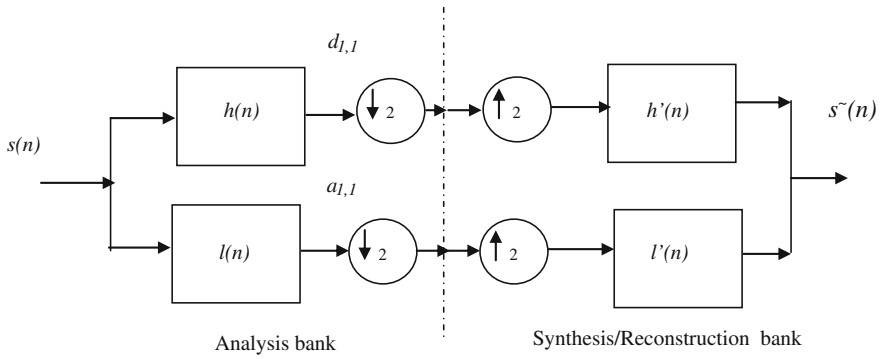

Fig. 3.1 Analysis and reconstruction of a signal using DWT through two-channel filter bank

$$\psi_{jk}(t) = 2^{j/2}\psi\left(2^j t - k\right), \tag{3.8}$$

where j and k are indices indicating scale and location of a particular wavelet.

Accordingly, the wavelet family is a collection of wavelet functions $\psi_{jk}(t)$ that are translated along the time axis t, then dilated by 2^j times and the new dilated wavelet is translated along the time again. The wavelets of a family share the same properties and their collection constitutes a complete basis. The basic wavelet function has necessarily to have local (or almost local) support in both a real dimension (time in case of speech signals) and frequency domain. Several kinds of wavelet functions have been developed and all of them have specific properties [3], as follows:

1. A wavelet function has a finite energy [4]

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty. \tag{3.9}$$

2. Similar condition must hold for $\psi(f)/f$ if $\psi(f)$ is the Fourier transform of the wavelet function and the wavelet function has zero mean $\psi(0) = 0$ or as follows:

$$\int_{0}^{\infty} \frac{|\psi(f)|^2}{f} df < \infty. \tag{3.10}$$

Another important property is that the wavelet function is compactly supported. The speed of convergence to 0, as the time t or the frequency goes to infinity, quantifies both time and frequency localizations. The symmetry is useful in avoiding dephasing. The number of vanishing moments for a wavelet function is

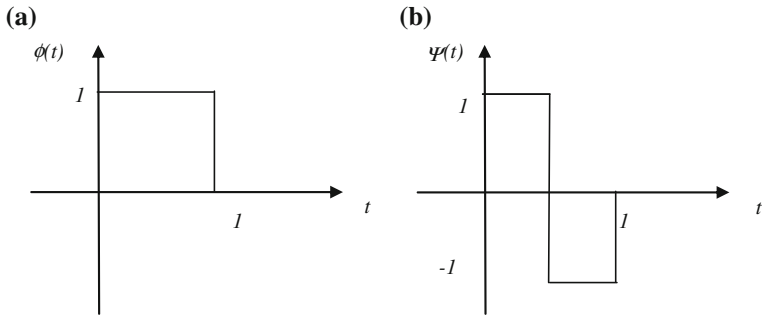

Fig. 3.2 Haar Scaling (**a**) and wavelet (**b**) functions

useful for compression purposes. The regularity can achieve smoothness of the reconstructed signal. Two other properties, namely the existence of a scaling function and orthogonality or biorthogonality, allow fast algorithm and space-saving coding.

There are a number of basis functions that can be used as the mother wavelet for wavelet transformation. Since the mother wavelet produces all wavelet functions used in the transformation through translation and scaling, it determines the characteristics of the resulting transform. Therefore, the appropriate mother wavelet should be chosen in order to use the wavelet analysis effectively for a specific application.

Figure 3.2 shows an example of the simplest wavelet, Haar. Haar wavelet is one of the oldest and simplest wavelets. The Haar scaling function acts as a low-pass filter through averaging effect on the signal, while its wavelet counterpart acts as a high-pass filter.

Daubechies wavelets are the most popular wavelets. They represent the foundations of wavelet signal processing and are used in numerous applications. The Haar, Daubechies, Symlets, and Coiflets are compactly supported orthogonal wavelets. These wavelets along with Meyer wavelets can provide a perfect reconstruction for signal. The Meyer, Morlet, and Mexican Hat wavelets are symmetric in shape [4]. The discrete form of a scale-function is the impulse response of a low-pass– filter, while the wavelet one is the impulse response of a high-pass filter.

3.5 Wavelet Packets

Wavelet packet basis consists of a set of multiscale functions derived from the shift and dilation of a basic wavelet function as in (3.8). The wavelet packet basis space is generated from the decomposition of both the low-pass filter function space and the corresponding basic high-pass filter function space. The conventional wavelet

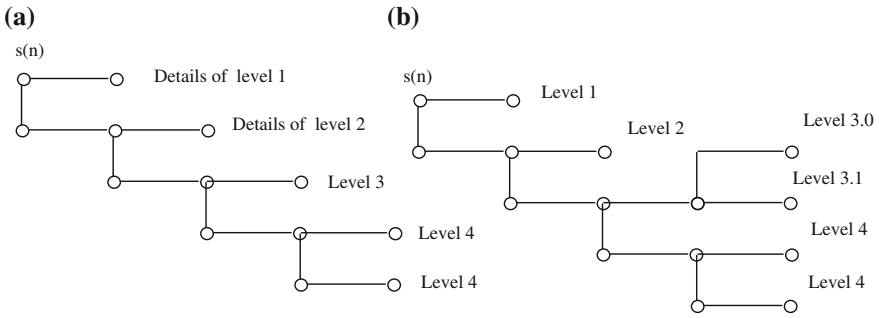

Fig. 3.3 A comparison between **a** conventional and **b** wavelet packet tree

basis space can be considered as a special case of the wavelet packet space when the decomposition takes place only in the low-pass filter function space [5]. Assuming that the discrete form of a scale-function is l(n) and the wavelet one is h(n), the wavelet packet basis can be expressed as

$$\begin{aligned} \psi_{01}(n) &= \sum_k \psi_{00}(k)\, l(2n-k) \\ \text{and} \qquad \psi_{11}(n) &= \sum_k \psi_{00}(k)\, h(2n-k) \end{aligned} \tag{3.11}$$

where $\psi_{00}(k)$ is the wavelet basis function with the finest time resolution. The functions in the next scale become coarser in time resolution and finer in spectral resolution through filtering and downsampling in (3.11). The same procedure can be applied recursively to the outputs of (3.11) into subsequent scales. In other words, a complete and orthogonal wavelet packet basis can be generated from a frequency decomposition tree which starts by using recursive two channel filtering and downsampling from an initial function with the finest time resolution. It can be shown that the decomposed functions at the outermost branches of the tree satisfy orthogonality and completeness for any decomposition tree and, thus, constitute a wavelet packet basis set [6].

Figure 3.3a and b show, respectively, an example of a conventional wavelet decomposition tree and a predefined wavelet packet (PWP) decomposition tree. The former always zooms in along the low-frequency branch while the latter zooms in along a preselected frequency.

The wavelet packet (WP) analysis provides much better frequency resolution than WT. Subbands with finer bandwidths across the whole spectrum can be attained using WP analysis (see Fig. 3.4).

First, the tree structure of WP decomposition can be chosen in a way to closely mimic the critical bands in a psychoacoustic model [7]. Several WP audio algorithms have successfully employed time-invariant WP tree structures that mimic the frequency resolution properties of ear's critical bands for perceptual quality assessment of speech [8] and [9].

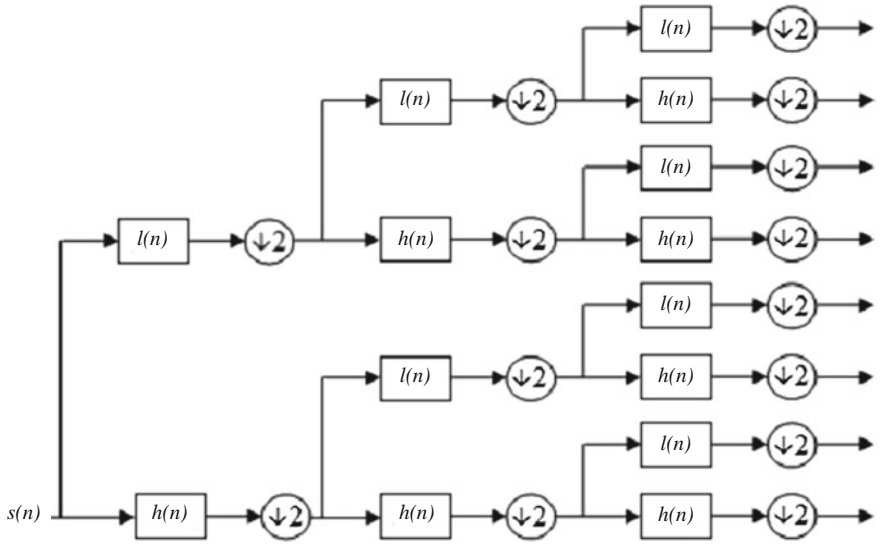

Fig. 3.4 Wavelet packet decomposition over 3 levels. *l[n]* is the low-pass approximation coefficients, *h[n]* is the high-pass detail coefficients

In order to achieve a critical band resolution using Fast Fourier Transform (FFT), it requires (N $\log_2$N) multiplications, and the whole process becomes computational intensive as *N* becomes larger where *N* is the number of samples per frame. WP can, directly, calculate the signal energy in the wavelet domain, and in turn, the complexity is greatly reduced [7].

References

1. T.K. Sarkar, C. Su, R. Adve, M. Salazar-Palma, L. Garcia Castillo, R.R. Boix, A tutorial on wavelets from an electrical engineering perspective, part II: the continuous case. IEEE Antennas Propag Mag **40**(6), 36–48 (1998)
2. V.T Pham, Wavelet analysis for robust speech processing and applications. Ph.D. Dissertation, 2007, http://theses.eurasip.org/media/theses/documents/pham-van-tuan-wavelet-analysis-for-robust-speech-processing-and-applications.pdf
3. R.J.E. Merry, *Wavelet Theory and Applications: A Literature Study* (Technische. Universiteit Eindhoven (Eindhoven), DCT 2005)
4. D. Sripathi, Efficient implementations of Discrete Wavelet Transforms using FPGAs, M.Sc. Thesis, Florida State University, 2003
5. R.R. Coifman, Y. Meyer, M.V. Wickerhauser, in *Wavelets and Their Applications*, ed. by M.B. Ruskai et al. Size Properties of Wavelet Packets, (Jones Bartlett, Boston, 1992)
6. M.V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software* (IEEE Press, New York, 1994)

7. W. Dobson, J. Yang, K. Smart, F. Guo, High quality low complexity scalable wavelet audio coding, in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP'97)*, Apr 1997, pp. 327–330
8. M. Hesham, M.H. Kamel, A unified mathematical analysis for acoustic wave propagation inside the vocal tract. J. Eng. Appl. Sci. **48**(6), 1099–1114 (2001)
9. A. Karmakar, A. Kumar , R.K. Patney, A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality. IEEE Trans. Audio Speech Lang. Process. **14**(6), 1912–1923 (2006)

Chapter 4
Speech Enhancement and Noise Suppression

Abstract Wavelet analysis has been widely used for noise suppression in signals. The multiresolution properties of wavelet analysis reflect the frequency resolution of the human ear. The wavelet transform (WT) can be adapted to distinguish noise in speech through its properties in the time and frequency domains.

Keywords Speech enhancement · Wavelet thresholding · Wavelet denoising

Speech enhancement aims to improve the quality and intelligibility of speech signals, as perceived by human hearing process. Figure 4.1 shows the main processes included during speech enhancement. Given that a noisy speech signal has an additive noise, the signal can be assumed as

$$s(n) = y(n) + n(n).$$

where $n(n)$ is the noise component in the signal and $y(n)$ is the original clean signal. After completing the enhancement process, it will be claimed that, the estimated clean signal $s^{\sim}(n)$ is approaching $y(n)$ in an optimal way as discussed below. As customarily expected, WT can be efficiently used as a spectral analysis tool.

In general, wavelets have been widely used for noise suppression in speech signals since the inception of wavelet analysis. The ideas of noise removing by WT were based on the singularity information analysis in [1] and the thresholding of the wavelet coefficients. Seminal works on signal denoising via wavelet thresholding or shrinkage of Donoho and Johnstone [2, 3] have shown that various wavelet thresholding schemes for denoising have near-optimal properties in the minimax sense. The wavelet denoising starts with WT using specific wavelet basis. Only few coefficients in the lower bands could be used for approximating the main features of the clean signal. Hence, by setting the smaller details to zero, up to a predetermined threshold value, we can reach a nearly optimal elimination of noise while preserving the important information of the clean signal.

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_4,

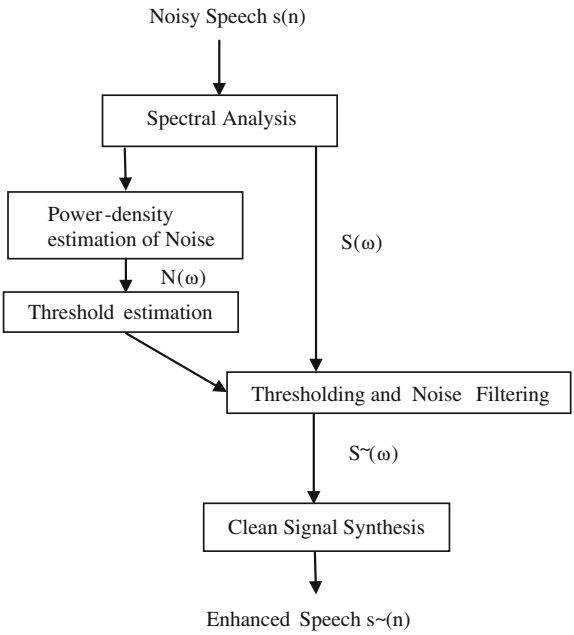

Fig. 4.1 Speech enhancement process

The thresholding (Shrink) approach comprises the following steps:

1. A forward WT of the observed data.
2. Thresholding the wavelet coefficients.
3. Inverse wavelet transform of the thresholded coefficients.

There are two thresholding methods frequently used. The soft threshold function takes the argument and shrinks it toward zero by the threshold. The other alternative is the hard threshold function which keeps the input if it is larger than the threshold; otherwise, it is set to zero. Hard thresholding maintains the scale of the signal but introduces ringing and artifacts after reconstruction due to a discontinuity in the wavelet coefficients. Soft thresholding eliminates this discontinuity resulting in smoother signals but slightly decreases the magnitude of the reconstructed signal [4]. The soft thresholding rule is preferred over hard thresholding for several reasons as discussed in [5]. Semisoft shrinking with selected threshold for unvoiced regions [6] and a smooth hard thresholding function based on μ-law in [6–8] were introduced to improve the shrinkage approach for denoising. The combination of soft and hard thresholding is applied to adapt with different properties of the speech signal in [9]. Wavelet thresholding methods are also integrated with other techniques such as the Teager energy operator and masked adaptive threshold in [10].

The estimation of the threshold seeks to minimize the maximum error over all possible samples of a signal. A universal threshold was proposed based on the

standard deviation of the signal. This threshold is shown to be asymptotically optimal in the minimax sense when employed as a hard threshold [4].

A more advanced strategy based on Stein's unbiased risk estimate (SURE) was proposed to get a threshold value [3]. A recent work in [11] considers undecimated wavelet transform (UWT) to avoid the drawback of the discrete WT (DWT) which is not shift invariant. In the literature, the method which employs a universal threshold may be called VisuShrink. SureShrink is a hybrid of the universal threshold and the SURE threshold when soft thresholding is used [5]. BayesShrink is an adaptive wavelet thresholding method proposed in [12] using a Bayesian estimate of the risk.

Another approach considers wavelet denoising with multitaper spectrum (MTS) estimation. However, the wavelet shrinkage approach has not been fully optimized for denoising the MTS of noisy speech signals. In [13], a two-stage wavelet denoising algorithm was proposed for estimating the speech power spectrum. The wavelet transform is applied to the periodogram of a noisy speech signal and the resulting wavelet coefficients are searched to indicate the approximate locations of the noise floor in the periodogram. The wavelet coefficients of the noise floor are then selectively removed in the *log MTS* of the noisy speech. The wavelet coefficients that remain are then used to reconstruct a denoised MTS. Simulation results outperform the traditional shrinking approaches and improve both the quality and intelligibility of the enhanced speech [13].

Other works consider WP coefficients instead of DWT ones in [14–16]. Recently, some works treat the wavelet packet tree in a perceptual manner [17], in which, the perceptual wavelet filterbank (PWF) is built to approximate the critical band responses of the human ear. Critical band wavelet decomposition is used with noise masking threshold in [18], and perceptual wavelet packet decomposition (PWPD) which simulates the critical bands of the psychoacoustic model is proposed in [15] and [19]. In [20], an adaptive threshold is statistically determined and applied to WP coefficients of noisy speech through a hard thresholding function. Several standard objective measures and subjective observations show that the proposed method outperforms recent state-of-the-art thresholding-based approaches from high- to low-level SNRs.

Alternatively, the wavelet transform (WT) has been used for blind adaptive filtering of speech signals from unknown colored noise when neither speech nor noise is separately accessible in [21].

References

1. S. Mallat, W. Hwang, Singularity detection and processing with wavelets. IEEE Trans. Inform. Theory **38**(2), 617–643 (1992)
2. D.L. Donoho, I.M. Johnstone, I.M, Ideal spatial adaptation via wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)
3. D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. **90**(432), 1200–1224 (1995)

4. V. Balakrishnan, N. Borges, L. Parchment, Wavelet denoising and speech enhancement, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, Spring, 2006
5. S.G. Chang, Y. Bin, M. Vetterl, Adaptive wavelet thresholding for image denoising and compression. Image Process. IEEE Trans. **9**(9), 1532–1546 (2000)
6. V.T. Pham, Wavelet analysis for robust speech processing and applications, Ph.D. Dissertation, 2007, http://theses.eurasip.org/media/theses/documents/pham-van-tuan-wavelet-analysis-for-robust-speech-processing-and-applications.pdf
7. H. Sheikhzadeh, H.R Abutalebi, An improved wavelet-based speech enhancement system, in *Proceedings of Eurospeech*, 2001, pp. 1855–1858
8. S. Chang, Y. Kwon, S. Yang, I. Kim, Speech enhancement for nonstationary noise environment by adaptive wavelet packet, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, vol. 1, 2002, pp. 561–564
9. A. Lallouani, M. Gabrea, C.S. Gargour, Wavelet based speech enhancement using two different threshold-based denoising algorithms, in *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, 2004, pp. 315–318
10. M. Bahoura, J. Rouat, Wavelet speech enhancement based on the teager energy operator. IEEE Sig. Process. Lett. **8**(1), 10–12 (2001)
11. M.A Hassanein, M. El-Barawy, N.P.A. Seif, M.T. Hanna, Trimmed thresholding with SURE for denoising signals, Circuits and Systems (MWSCAS), in *2012 IEEE 55th International Midwest Symposium,* 5–8 Aug 2012, pp. 1024–1027
12. G. Chang, B. Yu, M. Vetterli, Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. Image Process. **9**(9), 1532–1546 (2000)
13. D. Pak-Kong Lun, S. Tak-Wai, H. Tai-Chiu, Dominic K.C. Ho, Wavelet based speech presence probability estimator for speech enhancement. Digit. Sig. Process. **22**(6), 1161–1173 (2012)
14. S. Chang, Y. Kwon, S.-i. Yang, I.-j. Kim, Speech enhancement for non-stationary noise environment by adaptive wavelet packet, Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP'2002) **1**, I-561–I-564 (2002)
15. S.-H. Chen, J.-F. Wang, Speech enhancement using perceptual wavelet packet decomposition and teager energy operator. J. VLSI Sig. Process. Syst. Sig. Image Video Tech. **36**, 125–139 (2004)
16. Y. Ghanbari, M.R.K. Mollaei, A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets. Speech Commun. 48, 927–940 (2006)
17. Y. Shao, C.-H. Chang, Bayesian separation with sparsity promotion in perceptual wavelet domain for speech enhancement and hybrid speech recognition. IEEE Trans Syst Man Cybern Part A: Syst Hum 41(2):284–293
18. C. Lu, H.C Wang, Enhancement of single channel speech based on masking property and wavelet transform. Speech Commun. **41**(3), 409–427 (2003)
19. E. Jafer, A.E. Mahdi, Wavelet-based perceptual speech enhancement using adaptive threshold estimation, in *Proceedings of Eurospeech*, 2003, pp. 569–572
20. T.F. Sanam, C. Shahnaz, Noisy speech enhancement based on an adaptive threshold and a modified hard thresholding function in wavelet packet domain. Digit. Sig. Process. **23**(3), 941–951 (2013)
21. D. Veselinovic, D. Graupe, A wavelet transform approach to blind adaptive filtering of speech from unknown noises. IEEE Trans. Circuits Syst. II: Analog Digital Sig. Process. **50**(3), 150–154 (2003)

Chapter 5
Speech Quality Assessment

Abstract The wavelet packet analysis can be used to improve a perceptual-based objective speech quality measure. In this measure, the critical bands of auditory system can be approximated by a predefined wavelet packet (PWP) tree structure.

Keywords Multiresolution auditory model · Wavelet packet analysis

Subjective listening tests are accurate for speech quality assessment but these tests may be slow and expensive, while an objective measure is more economic to give an immediate and reliable estimate of the anticipated perceptual quality of speech. Recent objective measures employ perception-based assessment features. Objective measure of perceived speech quality has two essential components; first a perceptual transformation and distance measure. A perceptual transformation represents a speech signal as estimated by human hearing system for higher perceiving system. The distance measure estimates the perceived contrast between two speech signals. Most of the objective measures use STFT as a spectrum estimation tool which represents a part of the perceptual transformation of the quality measure. However, the loudness perception has nonlinear nature and nonuniform frequency resolution for a time varying speech signal [1]. Accordingly, wavelet analysis can replace spectrum estimation techniques which are used in such tests. Wavelet multiresolution analysis provides accurate localization in both time and frequency domains which can emulate the human auditory system operation. Moreover, wavelet analysis can reduce the computational effort associated with the mapping of speech signals into an auditory scale [2]. The WP analysis can provide two merits for an efficient objective quality measure of speech [3]. First, the tree structure of WP decomposition can be chosen in a way to closely mimic the critical bands in a psychoacoustic model [4]. Several algorithms have successfully employed time-invariant WP tree structures that mimic the frequency resolution properties of the ear's critical bands for perceptual quality assessment of speech [5] and [6]. In many works like [5] and [6], the WPT is used as a perceptual transformation in an objective quality measure of speech. A predefined path along the packet wavelet tree is proposed to approximate the critical bands of human hearing. Each tree leaf is a filter which gives a band energy extracted from the

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_5,

speech signal. These band features of each signal frame represent a features vector which is then included in distance measure computation. A quality judgment is then taken based on the calculated distance measures.

As a new cognition module of the perceptual quality measurement system, the Wavelet-based Bark Coherence Function (WBCF) computes coherence function with perceptually weighted speech after wavelet series expansion [7]. By using the WBCF, it is possible to alleviate the effects of variable delay of packet-based end-to-end system and linear distortion caused by the analog interface of the communication systems [7].

References

1. S. Voran, Objective estimation of perceived speech quality—part I: development of measuring normalizing block technique. IEEE Trans. Speech Audio Proces. **7**(4), 371–382 (1999)
2. B. Carnero, A. Drgajlo, Perceptual speech coding and enhancement using frame synchronized fast wavelet packet transform algorithms. IEEE Trans. Signal Process. **47**(6), 1622–1635 (1999)
3. M. Siafarikas, T. Ganchev, N. Fakotakis, Objective wavelet packet features for speaker verification, in *Proceedings of INTERSPEECH 2004—ICSLP*, Jeju, Korea, Oct 2004, pp. 2365–2368
4. W. Dobson, J. Yang, K. Smart, F. Guo, High quality low complexity scalable wavelet audio coding, in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP'97)*, Apr 1997, pp. 327–330
5. M. Hesham, A predefined wavelet packet for speech quality assessment. J. Eng. Appl. Sci. **53**(5), 637–652 (2006)
6. A. Karmakar, A. Kumar, R.K. Patney, A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality. IEEE Trans. Audio Speech Lang. Process **14**(6), 1912–1923 (2006)
7. S.-W. Park, Y.-C. Park, D. Youn, Speech quality measure for VoIP using wavelet based bark coherence function, in *Proceedings of INTERSPEECH 2001*, 2001, pp. 2491–2494

Chapter 6
Speech Recognition

Abstract Wavelet analysis can be used to improve the speech recognition performance through two approaches. In the first approach, it can be used as the back-end to remove noise and consequently the recognition process may perform better. In the second approach, wavelet-based features can be added to other successful features to improve recognition performance.

Keywords Multiresolution auditory model · Wavelet packet analysis · Bark-scale features · Speech recognition

Automatic-speech recognition (ASR) systems generally carry out some kind of classification/recognition based upon speech features which are usually obtained via time–frequency representations. Accordingly, the speech waveform is converted into feature vectors. Acoustic and linguistic models are then used with the features to recognize the content of an utterance. A common set of feature vectors are variants of spectral representation of speech signal obtained by Fourier or cepstral analysis.

While real-world applications require that speech recognition systems be robust to interfering noise, the performance of a speech recognition system drops dramatically when there is a mismatch between training and testing conditions. Many different approaches have been studied to decrease the effect of noise on the recognition [1]. Wavelet denoising can be applied as a preprocessing stage before feature extraction to compensate noise effects [2].

Dealing with enhancement and feature extraction for robust ASR, several parameterization methods which are based on the DWT and WP decomposition (WPD) have been proposed in [3]. More sophisticated shrinking functions with better characteristics than soft and hard thresholding are optimized for speech enhancement and speech recognition in [4].

The most widely used speech representation is based on the mel-frequency cepstral coefficients (MFCC) inspired from models of human perception; however, they provide limited robustness, as evidenced by the difficulty of state-of-the-art systems to adapt to noise and distortions [3]. Recent advances have been made with the introduction of wavelet-based representations, which have shown to

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_6,

improve the classification performance. The WT overcomes some of the limitations faced with other features since it can be used to analyze a speech signal directly into the critical bands defined by a psychoacoustic model. WT has become a frequently used method for improvement of speech recognition in recent years as evidenced by denoising theory and practice [3].

Eventually, many works employ wavelet features in an ASR while such features are computed based on critically sampled filter bands using WT or WP analysis [1, 5–7]. The usage of wavelet-based features extracted from the WPD leads to improvement of recognition rate compared with the well-known conventional MFCC features [8] and [9]. However, WT features are not time invariant. Time variant property decreases the rate of speech recognition systems. On the other hand, due to the nature of discrete wavelet, exact auditory bandwidth as in Mel scale cannot be achieved [6]. Another perspective for optimizing wavelet speech recognition is presented in [10] and [11], and it is based on frequency aspect of continuous wavelet transform (CWT). The features are chosen based on Bark scale, and it is known as Bark wavelet. The performance of ASR with features based on Bark wavelet analysis is discussed and compared in [10] and [11].

Despite that removal of noise can improve the speech recognition accuracy, errors in the estimated signal components can also obscure the recognition. To overcome this, a wavelet-based framework is realized in [12] by implementing speech enhancement preprocessing, feature extraction, and a hybrid speech recognizer in the wavelet domain. A Bayesian scheme is applied in a wavelet domain to separate the speech and noise components in an iterative speech enhancement algorithm. The denoised wavelet features are then fed to a classifier. The intrinsic limitation of the used classifier is overcome by augmenting it with a wavelet support vector machine (SVM). This hybrid and hierarchical design paradigm improves the recognition performance at a low SNR without causing a poorer performance at a high SNR [12].

The hidden Markov model (HMM) is one of the most widely used and successful classifiers for speech recognition. An HMM is a stochastic process that can estimate the probability of an observed sequence generated by an HMM for a specific speech unit. Alternatively, another approach is proposed in [13] for speech recognition based on wavelet network (WN). The proposed system is a hybrid classifier. It is based on neural network (NN) as a general model and the wavelets assume the role of activation function. The obtained results show that an ASR system based on WN is very competitive compared to HMM-based systems. The WN-based systems benefit from the ability of wavelet analysis to resist noise and the adaptability of NN on the other side. Another work employs a wavelet SVM (WSVM) which improves the accuracy of a WN recognizer due to its multiscale property and robustness [14].

Since a wavelet-based representation should be searched for each particular problem, a genetic algorithm is employed in [15]. The representation search is based on a non-orthogonal wavelet decomposition for phoneme classification. The results obtained for a set of Spanish phonemes show that the proposed genetic algorithm is able to find a representation that improves speech recognition results [15].

References

1. Z. Tufekci, J.N. Gowdy, S. Gurbuz, E. Patterson, Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition. Speech Commun. 48(10), 1294–1307 (2006)
2. O. Farooq, S. Datta, Wavelet-based denoising for robust feature extraction for speech recognition. Electron. Lett. 39(1), 163–165 (2003)
3. M. Gupta, A. Gilbert, Robust speech recognition using wavelet coefficient features, in *Automatic Speech Recognition and Understanding, IEEE Automatic Speech Recognition and Understanding Workshop 2001 (ASRU'01)*, Madonna di Campiglio, Italy, 2001, pp. 445–448
4. B. Kotnik, Z. Kacic, B. Horvat, The usage of wavelet packet transformation in automatic noisy speech recognition systems. Int. Conf. Comput. Tool **2**, 131–134 (2003)
5. Z. Xueying, J. Zhiping, Speech recognition based on auditory wavelet packet filter, in *Proceedings of 7th International Conference on Signal Processing, 2004 (ICSP '04)*, vol. 1, 2004, pp. 695–698
6. O. Farooq, S. Datta, Mel filter-like admissible wavelet packet structure for speech recognition. IEEE Sig. Process. Lett. **8**(7), 196–198 (2001)
7. J.N. Gowdy, Z. Tufekci, Mel-scaled discrete wavelet coefficients for speech recognition, in *Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP 2000)*, vol. 3, 2000, pp. 1351–1354
8. V.T. Pham, Wavelet analysis for robust speech processing and applications, Ph.D. Dissertation, 2007, http://theses.eurasip.org/media/theses/documents/pham-van-tuan-wavelet-analysis-for-robust-speech-processing-and-applications.pdf
9. S. Chang, Y. Kwon, S. Yang, I. Kim, Speech enhancement for nonstationary noise environment by adaptive wavelet packet, in *Proceeedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, vol. 1, 2002, pp. 561–564
10. H.R. Tohidypour, S.A. Seyyedsalehi, H. Behbood, Comparison between wavelet packet transform, Bark Wavelet & MFCC for robust speech recognition tasks, in *Proceedings of 2nd International Conference on Industrial Mechatronics and Automation (ICIMA2010)*, vol. 2, 2010, pp. 329–332
11. Z. Jie; L. Guo-Liang; Z. Yu-zheng, L. Xiao-Ying, A novel noise-robust speech recognition system based on adaptively enhanced bark wavelet MFCC, in *Proceedings of Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009 (FSKD '09)*, vol. 4, 2009, pp. 443–447
12. Y. Shao, C.H. Chang, A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter bank modeling of human auditory system. IEEE Trans. Syst. Man Cybern. Part B Cybern. **37**(4), 877–889 (2007)
13. R. Ejbali, M. Zaied, C. Ben Amar, Wavelet network for recognition system of Arabic word. Int. J. Speech Technol. **13**(3), 163–174 (2010)
14. Y. Shao, C.-H. Chang, Bayesian separation with sparsity promotion in perceptual wavelet domain for speech enhancement and hybrid speech recognition. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **41**(2), 284–293 (2011)
15. L.D. Vignolo, D.H. Milone, H.L. Rufiner, Genetic wavelet packets for speech recognition. Expert Syst. Appl. 40(6), 2350–2359 (2013)

Chapter 7
Emotion Recognition from Speech

Abstract Like speech recognition, emotion recognition can benefit from the merits of wavelet analysis in feature extraction or mapping functions.

Keywords Wavelet analysis · Wavelet packet analysis · Emotion recognition · Neural network

The emotion detection from human speech has a wide variety of applications that benefit from such technology. Emotion recognition from speech can be viewed as a classification task between speech uttered by a human under different emotional conditions, like happiness or anger and so on. A survey on speech emotion classification [1] cited the work in [2] which addressed the use of wavelets in two classifiers: one for a video part and another for the audio part of an emotional database. Features were extracted from the video data using multiresolution analysis based on the DWT. The dimensionality of the obtained wavelet coefficients vector is then reduced. The same was repeated for the audio data. The two feature sets are then combined. The fusion algorithm was applied to another database which contains the following emotions: happiness, sadness, anger, surprise, fear, and dislike. The recognition accuracies were 98.3 % for female speakers and 95 % for male speakers.

In [3] two different WP filterbank structures were proposed based on Bark scale and equivalent rectangular bandwidth (ERB) scale for multistyle classification of speech under stress. Linear Discriminant analysis (LDA)-based classifier is then used for emotion recognition. The experimental results reveal a classification accuracy of more than 90 %.

Another work in [4] DWT coefficients were used as features while an NN was used for pattern classification. Daubechies type of mother wavelet was used for DWT. Overall recognition accuracies of 72.05, 66.05, and 71.25 % were obtained for male, female, and combined male and female databases, respectively.

Alternatively, features relevant to energy, speech rate, pitch, and formant are extracted from speech signals in [5]. WN is used as the classifier for five emotions, including anger, calmness, happiness, sadness, and boredom. Compared to the

M. H. Farouk, *Application of Wavelets in Speech Processing*, SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_7,

traditional back-propagation (BP) NN, the results of experiments show that the WN has faster convergence speed and higher recognition rate.

Different features were extracted in [6] to identify the parameters responsible for emotion. WPT is proved to be emotion specific. The experiments include the comparison of WPT coefficients with a threshold in different bands. In another experiment, energy ratios based on WPT were compared in different bands. The results are then compared to the conventional method using MFCC. Based on WPT features, a model is also proposed for emotion conversion, namely neutral to angry and neutral to happy emotion.

References

1. M. El-Ayadi, M.S. Kamel, F. Karray F, Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
2. H. Go, K. Kwak, D. Lee, M. Chun, Emotion recognition from the facial image and speech signal. In: Proceedings of SICE annual conference 2003, vol. 3, pp. 2890–2895 (2003)
3. N.A. Johari, M. Hariharan, A. Saidatul, S. Yaacob, Multistyle classification of speech under stress using wavelet packet energy and entropy features. in *Proceedings of IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT 2011)*, 2011, pp. 74–78
4. F. Shah, A.R. Sukumar, A.P. Babu, Discrete Wavelet Transforms and artificial neural networks for speech emotion recognition. Int. J. Comput. Theory Eng. **2**(3), 1793–8201 (2010)
5. Y. Huang, G. Zhang, X. Xu, Speech emotion recognition research based on wavelet neural network for robot pet, in *Proceedings of the Intelligent Computing 5th International Conference on Emerging Intelligent Computing Technology and Applications (ICIC'09)*, ed. by De-Shuang Huang, Kang-Hyun Jo, Hong-Hee Lee, Hee-Jun Kang, and Vitoantonio Bevilacqua (Eds.), Springer-Verlag, Berlin, Heidelberg, 2009, pp. 993–1000
6. V.N. Degaonkar, S.D. Apte, Emotion modeling from speech signal based on wavelet packet transform. Int. J. Speech Technol. **16**(1), 1–5 (2013)

Chapter 8
Speaker Recognition

Abstract MFCC features are widely used in speaker recognition. However, MFCC is not suitable for identifying a speaker since they should be located in high frequency regions, while the Mel scale gets coarser in the higher frequency bands. The speaker individual information, which is nonuniformly distributed in the high frequencies, is equally important for speaker recognition; accordingly, wavelet-based features are more appropriate than MFCC.

Keywords Wavelet analysis · Speaker identification · Speaker recognition · Speaker verification

Speaker recognition deals with the problem of identifying the talker from others, while speaker verification is concerned with ascertaining the identity of a speaker. Both techniques have many useful applications for security. Speaker recognition systems are composed of a feature extraction stage and a classification one. Feature extraction is concerned with extracting speaker's characteristics while avoiding any sources of adverse variability. The resulting feature vector makes use of information from all spectrum bands, and therefore, any inaccuracy of representation and any distortion induced to any part of the spectrum are spread to all features.

The recognition accuracy of current speaker recognition systems under controlled conditions is high. However, in practical situations many negative factors are encountered, including mismatched handsets for training and testing, limited training data, unbalanced text, background noise, and noncooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation, and score normalization methods are necessary.

MFCC features are widely used in many speaker recognition systems, however, one of the weaknesses of MFCC is that the speech signal is assumed to be stationary within the given time frame [1]. Alternatively, the use of wavelet filterbanks is a better choice in many recent works for speaker recognition [2]. The classification stage identifies the feature vector with certain class and is based

M. H. Farouk, *Application of Wavelets in Speech Processing*, SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_8,

on the probability density function of the acoustic vectors which is seriously confused in case of impaired features.

A work in [3] used WP-based sets as speech features. Comparative experimental results confirm the assertion that the WP-based features outperform MFCC, as well as other wavelet features, on the task of speaker verification.

Another work in [4] describes a speaker verification that includes what is called the wavelet octave coefficients of residues (WOCOR) as features. The proposed features capture the spectro-temporal source excitation characteristics embedded in the linear predictive residual signal. WOCOR is used to supplement the conventional MFCC features for speaker verification. Speaker verification experiments which are carried in that work, reveal an equal error rate (EER) of 7.67 % using the wavelet-based method, in comparison to 9.30 % of the conventional MFCC-based system on NIST-database.

Another trial is presented in [5] in which a wavelet-based filter structure is fine-tuned to some of the frequency bands which are more important for speaker discrimination. This structure does not follow the human auditory band structure. The study shows improved identification performance compared to other commonly used Mel scale-based filter structures using wavelets and proves the need for a filter structure to extract speaker-specific features.

A hybrid technique is proposed in [1] for extracting the features by combining the AM–FM modulation/demodulation approach with WT analysis. Features are extracted from the envelope of the signal and then passed through wavelet filterbank. From the results it is observed that the features extracted from envelope of the signal are more robust against noise. Experimental results in [1] show that the proposed hybrid method gives better efficiency for speaker identification as compared to AM-based method, MFCC, WP, and wavelet filterbanks.

In the manner of hybrid classifiers, Discrete wavelet-Fourier transform (DWFT) analysis can help in the case of analysis of quasi-harmonic signals, like speech signals. The DWFT can reveal some spectrum irregularities which are not directly visible in either the wavelet or Fourier spectrum. These irregularities together with particular behavior in the time domain determine specific properties of a sound. Consequently, a hybrid technique is presented in [6] combining DWFT and MFCC-based classifiers for classifying the sounds into voiced and unvoiced (V/U) segments. The results of such hybrid system outperform that based on MFCC features.

References

1. V. Tiwari, J. Singhai, Wavelet based noise robust features for speaker recognition. Signal Process.: Int. J. (SPIJ) **5**(2), 52–64 (2011)
2. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: From features to supervectors. Speech Commun. **52**(1), 12–40 (2010)

3. T. Ganchev, M. Siafarikas, N. Fakotakis, in *Speaker Verification Based on Wavelet Packets, Lecture Notes in Computer Science*, vol. LNAI 3206/2004 (Springe, Heidelberg, 2004), pp. 299–306
4. N. Zheng, T. Lee, P. Ching, Integration of complementary acoustic features for speaker recognition. IEEE Signal Process. Lett. **14**(3), 181–184 (2007)
5. S.M. Deshpande, R.S. Holambe, Speaker identification using admissible wavelet packet based decomposition. Int. J. Inf. Commun. Eng. **6**(1), 20–23 (2010)
6. B. Ziółko, W. Kozłowski, M. Ziółko, R. Samborski, D. Sierra, J. Gałka, Hybrid wavelet-fourier-HMM speaker recognition. Int. J. Hybrid Inf. Technol. **4**(4), 25–42 (2011)

Chapter 9
Spectral Analysis of Speech Signal and Pitch Estimation

Abstract Wavelet transform (WT) provides a way to explore the characteristics of nonstationary speech signals. Both time and frequency analysis can be conducted by WT. The tree structure of WP analysis can be customized to the critical bands of human hearing giving better spectral estimation for speech signal than other methods. Wavelet-based pitch estimation assumes that the glottis closures are correlated with the maxima in the adjacent scales of the WT. This approach ensures more accurate estimation of pitch period.

Keywords Wavelet analysis · Spectral analysis · Pitch estimation · Pitch detection · Fundamental frequency estimation

Spectral analysis of speech signal takes many forms in the context of speech processing algorithms. FFT, MFCC, linear-predictive code (LPC), and Cepstral analysis are examples of such forms. Multiresolution analysis based on the wavelet theory permits the introduction of concepts of signal filtering with different bandwidths or frequency resolutions. Since the speech signal is nonstationary, WT provides a framework to obtain more elegant spectral analysis through partitioning of sound in time according to its spectral properties [1].

Wavelet-spectral analysis has been used to measure the nonlinear content of a speech signal through wavelet scalogram [2]. WT can decompose the speech signal into a number of new signals, each with different degrees of resolution both in time and frequency. The tree structure of WP analysis can be customized to mimic the critical bands of human hearing giving better spectral estimation for speech signal than other methods [3–5].

The DWT is used for spectral analysis and to create a segmentation profile for a speech signal in [1] and the efficiency of the segmentation results are tested against the hand-annotated speech corpus.

Fast variation of a speech signal in both time and frequency domains can be detected by inspecting the decomposed coefficients of WT, and accordingly, obtaining more insight into signal spectrum. Moreover, abrupt changes of speech can be tracked by wavelet analysis, therefore, WT has been used successfully for pitch estimation of a speech signal and V/U classification [6]. Another work

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_9,

applied a data fusion method including wavelet features in [7] for both pitch estimation and speech segmentation.

It is difficult to estimate the pitch period due to the inherent large variability [6]. Traditional algorithms for pitch detection can be classified into two types, spectral-domain (nonevent)-based and time-domain (event)-based pitch detectors. The spectral-based pitch detectors, such as the autocorrelation and the cepstrum methods, estimate the average pitch period over a fixed-length window of a speech signal. The time-based pitch detectors estimate the pitch period by measuring the period between two successive instants of glottal closures. Pitch period can, in some sense, be related to finding the local maximum of its wavelet representation [6]. Several wavelet-based pitch determination algorithms have been presented in [8–11].

Pitch estimation approach which assumes that the glottal closures are correlated with the maxima in the adjacent scales of the WT is often prone to error especially in the case of noisy signals. In [12], an optimization scheme is proposed in the wavelet framework for pitch detection using a multipulse excitation model for the speech signal. Experimental results on both clean and noisy conditions show that the proposed optimization works better than the widely used heuristic approach for maxima detection.

In [13], WT is applied to the excitation part in the cepstrum of a speech signal. The local maximum is then searched within WT coefficients in order to extract the global peak index which represents the pitch. Using this method, the experimental results have been proved as more efficient than other classical approaches for pitch estimation. The local maxima is also searched in [14] through WT coefficients after pre-filtering for pitch detection. Such an approach exhibits superior performance compared to other wavelet methods in both clean and noisy environments.

Another event-based detection of pitch is developed in [15] and the local maximum is also searched through dyadic WT coefficients. The candidate pitches according to such maxima at each scale are then averaged. An optimal scale is chosen at minimum average of consecutive pitches and the optimal value of pitch period is estimated at such optimal scale. Experiments using this method show superior performance in comparison with other event-based detectors and classical ones that use the autocorrelation and cepstrum methods.

The method of modified higher order moment can replace the autocorrelation function which is used traditionally in pitch detection. In this method the average of the product of K samples are taken instead of two samples. It was shown that the modified higher order moment method is a better pitch estimator compared to traditional pitch detection techniques [16]. In [16], the higher order moment is applied to the transformed signal by dyadic WT and then, the local maxima are obtained. Pitch detection is achieved through searching within such maxima. The results in [16] show an improvement over conventional dyadic WT method even after adding noise.

A different technique in [17] uses the DWT to extract the wavelet parameters of noisy speech in the fundamental frequency band, and then performs variance analysis to generate a variance statistical distribution function of the wavelet parameters. The peak detection approach is then applied to extract the pitch period.

Due to the combination of the contribution of both WT and variance-analysis (VA), the proposed method can be effectively applied to pitch period estimation of speech with noise. The simulation results show that the proposed method can give a superior accuracy and robustness against noise relative to some of the existing methods from high to very low SNR levels.

References

1. J. Galka, M. Ziolko, Wavelets in speech segmentation, in *The 14th IEEE Mediterranean Electrotechnical Conference, MELECON2008*, 2008, pp. 876–879
2. M. Hesham, Wavelet-scalogram based study of non-periodicity in speech signals as a complementary measure of chaotic content. Int. J. Speech Technol. **16**(3), 353–361 (2013)
3. M. Hesham, A predefined wavelet packet for speech quality assessment. J. Eng. Appl. Sci. **53**(5), 637–652 (2006)
4. A. Karmakar, A. Kumar, R.K. Patney, A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality. IEEE Trans. Audio Speech Lang. Process. **14**(6), 1912–1923 (2006)
5. W. Dobson, J. Yang, K. Smart, F. Guo, High quality low complexity scalable wavelet audio coding, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Apr 1997, pp. 327–330
6. J.F. Wang, S.H. Chen, J.S. Shyuu, Wavelet transforms for speech signal processing. J. Chin. Inst. Eng. **22**(5), 549–560 (1999)
7. D. Charalampidis, V.B. Kura, Novel wavelet-based pitch estimation and segmentation of non-stationary speech, in *8th International Conference on Information Fusion*, , vol. 2, 2005, pp. 1383–1387
8. M. Obaidat, C. Lee, B. Sadoun, D. Neslon, Estimation of pitch period of speech signal using a new dyadic wavelet transform. J. Inform. Sci. **119**, 21–39 (1999)
9. M. Obaidat, A. Bradzik, B. Sadoun, A performance evaluation study of four wavelet algorithms for the pitch period estimation of speech signals. J. Inform. Sci. **112**, 213–221 (1998)
10. E. Ercelebi, Second generation wavelet transform based pitch period estimation and voiced/unvoiced decision for speech signals. Appl. Acoust. **64**, 25–41 (2003)
11. S. Kadambe, F. Boudreaux-Bartels, Application of the wavelet transform for pitch detection of speech signals. IEEE Trans. Inf. Theory **38**(2), 917–924 (1992)
12. P. Ghosh, A. Ortega, S. Narayanan, Pitch period estimation using multipulse model and wavelet transform, in *Proceedings of INTERSPEECH, ICSLP2007*, Antwerp, Belgium, Aug 2007, pp. 2761–2764
13. F. Bahja, E.-H. Ibn Elhaj, J. Di Martino, On the use of wavelets and cepstrum excitation for pitch determination in real-time, in *International Conference on Multimedia Computing and Systems (ICMCS)*, Tangier, Morocco, 2012, pp. 150–153
14. C. Runshen, Z. Yaoting, S. Shaoqiang, A modified pitch detection method based on wavelet transform, in *Second International Conference on Multimedia and Information Technology (MMIT), 2010*, Kaifeng, China, vol. 2, 2010, pp. 246–249
15. S. Bing, G. Chuan-qing, J. Zhang, A New Pitch Detection Algorithm Based on Wavelet Transform. J. Shanghai Univ. (English Edition) **9**(4), 309–313 (2005)
16. J. Choupan, S. Ghorshi, M. Mortazavi, F. Sepehrband, Pitch extraction using dyadic wavelet transform and modified higher order moment, in *12th IEEE International Conference on Communication Technology (ICCT), 2010*, Nanjing, China, 2010, pp. 833–836
17. X. Wei, L. Zhao, Q. Zhang, J. Dong, Robust pitch estimation using a wavelet variance analysis model. Signal Process. **89**(6), 1216–1223 (2009)

Chapter 10
Speech Coding, Synthesis, and Compression

Abstract WT-based coding allows for the control of frequency resolution to closely match the response of the human auditory system. The inherent shaping of the wavelet synthesis filter and a controlled bit allocation to the wavelet coefficients help to minimize the perceptually significant noise due to the quantization error in the residual. Experimental results show that WT-coders deliver superior quality to some audio standards when operating at the same bit rate and comparable quality to other codecs at lower bit rates. As a result, transform compression using WT can provide an efficient and flexible scheme for audio compression.

Keywords Audio coding · Subband coding · Wavelet transform coding · Warped filter

Speech coding is a major issue in the area of digital processing of speech signals. Speech coding is the act of transforming the speech signal to a more compact form, which can then be used in signal compression. Speech compression aims at reducing the required bandwidth for communication or storage size. Therefore, there is a need to code and compress speech signals. Speech compression is required in long-distance communication, high quality speech storage, and message encryption. Speech coding of lossy type maintains that the perceived quality is kept comparable to the original.

Several techniques of speech coding such as LPC, Waveform Coding, and Subband Coding have been used since many years for speech synthesis, analysis, and compression. However, there are many trials to apply the wavelet analysis for speech coding and compression. In fact, various methods have been developed based on wavelet or wavelet packet analysis for compressing speech signals as in [1–5]. These works also include analysis-by-synthesis for speech signals through wavelets [4, 5]. Most of such trials exploit the sparseness of speech signal representation in the wavelet domain. WT can concentrate speech information into a few neighboring coefficients. Therefore, such coefficients will either be zero or will have negligible magnitudes [6, 7]. In [7], the use of WT ensures lower complexity at arbitrary rates with acceptable quality if compared to other traditional techniques.

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_10,

The WT has been also used in compressing residuals of linear-prediction (LP) analysis and achieved good performance in Deriche and Ning, Shimizu et al. [8, 9].

A major issue in the design of real-time wavelet-based speech coder is choosing optimal wavelets for compression. The performance of the different wavelet families on speech compression is evaluated and compared in Joseph and Anto [10].

References

1. E.B. Fgee, W.J. Philips, W. Robertson, Comparing audio compression using wavelet with other audio compression schemes, in *Proceedings IEEE Electrical and Computer Engineering*, vol. 2, (1999), pp. 698–701
2. S. Dusan, J.L. Flanagan, A. Karve, M. Balaraman, Speech compression using polynomial approximation. IEEE Trans. Audio Speech Lang. Process **15**(2), 387–397 (2007)
3. S.M. Joseph, Spoken digit compression using wavelet packet. In *IEEE International Conference on Signal and Image Processing (ICSIP-2010)*, (2010), pp. 255–259
4. S. Moriai, I. Hanazaki, Application of the wavelet transform to the low-bit-rate speech coding system. Electr. Eng. Jpn. **148**(3), 62–71 (2004)
5. A. Kumar, G.K. Singh, G. Rajesh, K. Ranjeet, The optimized wavelet filters for speech compression. Int. J. Speech Technol. **16**(2), 171–179 (2013)
6. S.M. Joseph, P. Babu Anto, Speech compression using wavelet transform, in *International Conference on Recent Trends in Information Technology (ICRTIT2011)*, (2011), pp. 754–758
7. M. Abo-Zahhad, A. Al-Smadi, S.M. Ahmed, High-quality low-complexity wavelet-based compression algorithm for audio signals. Electr. Eng. **86**(4), 219–227 (2004)
8. M. Deriche, D. Ning, A novel audio coding scheme using warped linear prediction model and the discrete wavelet transform. IEEE Trans. Audio Speech Lang. Process **14**(6), 2039–2048 (2006)
9. T. Shimizu, M. Kimoto, H. Yoshimura, N. Isu, K. Sugata, A method of coding lsp residual signals using wavelets for speech synthesis. Electr. Eng. Jpn. **148**(3), 54–60 (2004)
10. S. Joseph, P. Anto, in *The optimal wavelet for speech compression. Communications in Computer and Information Science*, eds. by A. Abraham et al. ACC 2011, Part III, volume CCIS 192 (Springer, 2011), pp. 406–414

Chapter 11
Speech Detection and Separation

Abstract Many methods which are used for speech detection usually fail when signal-to-noise ratio (SNR) is low. The wavelet analysis has properties which can help in separating speech from noise. Many works report a better detection performance using wavelet analysis than other techniques.

Keywords Speech detection · Signal separation · Time–frequency analysis · Zero-crossing rate · Time energy

Speech detection means the localization of speech part in a signal containing other signals or noise. Many parameters are employed for speech detection and separation from other superimposed signals. Some such parameters are the time energy (the magnitude in time domain), zero-crossing rate (ZCR), cepstral coefficients, pitch information, and the time–frequency parameter. Detection in the presence of variable-level noise is more challenging than in the presence of impulse noise or fixed-level noise. A robust speech detection method in the presence of different types of noises with various levels is necessary for many practical applications [1].

Depending on the characteristics of speech, a variety of parameters have been proposed for speech detection. They include the time energy (the magnitude in time domain), ZCR, cepstral coefficient, pitch information, and the time–frequency parameter. These parameters usually fail to detect speech when signal-to-noise ratio (SNR) is low [1]. WT can reduce the influences of different types of noise at different levels.

The wavelet energy is used in [1, 2] to detect speech activity through a recurrent fuzzy NN. Results are obtained in these works with different types of noises and various SNRs. Comparing the results with other robust detection methods have verified the robust performance of such methods.

In a different work [3], WT is used to implement a voice activity detector (VAD) for European Telecommunication Standards Institution (ETSI) adaptive multi-rate (AMR) narrow-band (NB) *(ETSI AMR-NB)* and wide-band (WB) speech codecs. The original IIR filter bank and pitch/tone detector in the codec are reimplemented, respectively, via the wavelet filter bank and the wavelet-based pitch/tone detection algorithm. The wavelet filter bank can divide input speech

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_11,

signal into several frequency bands. The background noise level can also be estimated in each subband by using the wavelet denoising method. The wavelet filter bank is also derived to detect correlated complex signals like music. An SVM is then used to train the VAD decision rule involving the subband power, noise level, pitch period, tone flag, and complex signals warning flag of input speech signals. Experimental results show that the proposed algorithm in [3] gives considerable VAD performance compared to the standard one. SVM is used, again, in [4] for building a VAD based on MFCC of multiresolution spectrum via WT. Experiments on this VAD achieve robustness in all SNRs than VAD of G.729b, and its computational time delay satisfies the needs of real-time transmission on G.729b.

WP is also used in [5] to propose features across the frequency and time for VAD. The feature extraction is based on observations of the angles between the vectors in feature space. These WPT-based features also help to distinctly discriminate the voiced, unvoiced, and transient components of speech. Experimental results show that the proposed WPT-based approach in [5] is sufficiently robust such that it can extract the speech activity under poor SNR conditions and is also insensitive to variable level of noise.

References

1. C. Juang, C. Cheng, T. Chen, Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network. Expert Syst. Appl. **36**(1), 321–332 (2009)
2. C.C. Tu, C. Juang, Recurrent type-2 fuzzy neural network using Haar wavelet energy and entropy features for speech detection in noisy environments. Expert Syst. Appl. **39**(3), 2479–2488 (2012)
3. S.H. Chen, R. Guido, T.K. Truong, Y. Chang, Improved voice activity detection algorithm using wavelet and support vector machine. Comput. Speech Lang. **24**(3), 531–543 (2010)
4. W. Xue, S. Du, C. Fang, Y. Ye, Voice activity detection using wavelet-based multiresolution spectrum and support vector machines and audio mixing algorithm. in *Computer Vision in Human-Computer Interaction. Lecture Notes in Computer Science*, vol. 3979 (Springer, 2006), pp. 78–88
5. M. Eshaghi, M. Mollaei, Voice activity detection based on using wavelet packet. Digit. Sig. Proc. **20**(4), 1102–1115 (2010)

Chapter 12
Steganography and Security of Speech Signal

Abstract Perfect reconstruction of wavelet filter banks helps in retrieving a hidden signal. In wavelet domain different techniques are applied on the wavelet coefficients to increase the hiding capacity and perceptual transparency. In general, steganography in wavelet domain shows high hiding capacity and transparency.

Keywords Data hiding · Watermarking · Steganography · Encryption · Secure communication of speech

There are two classical approaches for secure transmission of a speech signal: encrypting and hiding the signal (steganography) [1]. In the first case, the speech signal is scrambled through a key; in the second case, the speech signal is hidden into a host signal and the resulting signal should be highly similar to the host one. In steganography, no one apart from the owners of the hidden message realizes its existence. The message data is hidden in an unremarkable media so that it becomes unremarkable. Secret signal which may be called stegano signal is hidden in the redundant portion of a cover medium [2]. Steganography, cryptography, and watermarking perform the same purpose in securing information. Alternatively, the secret message is converted into a different form in cryptography, while digital watermarking hides information in a carrier signal.

Commonly used techniques for audio steganography are either in time domain or transform domain. WT as a transform domain is preferred because of its multiresolution properties that provide access to both most significant parts and details of a spectrum.

In [3], the audio signal is transformed into wavelet domain. The wavelet coefficients are then scaled using the maximum value inside all subbands. The data bits of stegano signal are embedded in the least significant bit (LSB) portion of wavelet coefficients. At the receiver side, the secret bits are retrieved from these LSB portions. The stegano signal can be reconstructed because of perfect reconstruction properties of wavelet-filter banks. The use of Haar wavelet, in [3], shows large advantage in terms of hiding capacity and transparency compared to other methods.

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_12,

Mixed-Excitation Linear Prediction (MELP) algorithm is used in [4] to code the stegano speech into binary parameter bits. The cover speech is then divided into voiced and unvoiced frames using auditory WT. For voice frames, the auditory WT was used to detect pitch, and the pitch is utilized to localize the embedding position in cover medium. The information hiding procedure is completed by modifying relevant wavelet coefficients. Based on the same pitch detection method, the embedding position is found and the hiding bit is recovered. The stegano speech can be retrieved after MELP decoding. The experiments show that the method is robust to many attacks such as compression, filter, and so on.

In [5], Frequency Masking in human auditory system is used in hiding a speech signal into another audio content through sorting of the wavelet coefficients of the secret messages and indirect LSB substitution. This approach proves to have a hiding capacity significantly higher than the Spread and Shift Spectrum algorithms and additionally a statistical transparency higher than other mechanisms. Moreover, the transparency is not dependent on the host signal chosen because the wavelet sorting guarantees the adaptation of the secret message to the host signal. Similarly in [1], the masking property through wavelet coefficients helps in achieving real-time hiding of secret speech into another speech signal.

A different approach is presented in [6], in which the high-frequency components of a secret speech are separated from the low-frequency components using the DWT. In a second phase, FT is applied to the high-frequency components to get deeper spectral components. Finally, low-pass spectral components of stegano speech are hidden in the low-amplitude high-frequency regions of the cover speech signal. This method allows hiding a large amount of stegano information with slight degradation in quality.

In [7], chaotic logistic mapping is used to encrypt the speech signal. In order to hide the characteristics of information signal, the chaotic mapping is applied to the WT of secrete speech. The results show that the chaotic encrypting system can improve the security and can resist the decipher analysis with less complexity.

References

1. D.M. Ballesteros, J.M. Moreno (2013) Real-time, speech-in-speech hiding scheme based on least significant bit substitution and adaptive key. Comput. Electr. Eng. **39**(4), 1192–1203 (2013)
2. J. Antony, C. Sobin, A. Sherly, Audio steganography in wavelet domain—a survey. Int. J. Comput. Appl. **52**(13), 33–37 (2012)
3. N. Cvejic, T. Seppanen, A wavelet domain LSB insertion algorithm for high capacity audio steganography, in *Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop, 2002 and the IEEE 2nd Signal Processing Education Workshop* (2002)
4. L. Shen, X. Li, H. Wang, R. Zhang, in *Speech hiding based on auditory wavelet*, in *Proceedings of International Conference on Computational Science and Its Applications—ICCSA 2004*, Part IV (Assisi, 2004), pp. 414–420
5. D.M. Ballesteros, J.M. Moreno, Highly transparent steganography model of speech signals using efficient wavelet masking. Expert Syst. Appl. **39**(10), 9141–9149 (2012)

6. R. Siwar, G. Driss, S. Sid-Ahmed, H. Habib, Speech steganography using wavelet and fourier transforms. EURASIP J. Audio Speech Music Process. **2012**(20), 1–14 (2012)
7. D. Que, L. Lu, H. Wang, Y. Ding, A digital voice secure communication system based on logistic-mapping in wavelet domain chaotic modulation. in *7th International Conference on Signal Processing, 2004, ICSP '04*, vol. 3 (2004), pp. 2397–2400

Chapter 13
Clinical Diagnosis and Assessment of Speech Disorders

Abstract WT coefficients for normal voiced signal have remarkable differences compared to pathological ones. Accordingly, WT is successfully used as a non-invasive method to diagnose vocal pathologies.

Keywords Voice pathology detection and discrimination · Clinical diagnosis of speech disorders · Voice disorders identification

Clinical diagnostic symptoms of pathological speech include hypernasality, breathiness, diplophonia, audible nasal emissions, short phrases, pitch breaks, monopitch, monoloudness, reduced loudness, harshness, low pitch, slow rate, short rushes of speech and reduced stress, hypotonicity, atrophy, facial myokymia, fasciculation, occasional nasal regurgitation, dysphagia, and drooling. Impairments in the vocal folds can result in disorders in expressing words. Most of these disorders can be analyzed and graded from audio recordings of speech [1, 2]. In such pathological speech signal, strong high-frequency components can be detected as abnormal noise through spectral analysis and may be separated within a wavelet spectrum [3]. However, inspecting WT coefficients for normal voiced signal and pathological subjects show that amplitude of wavelet coefficients for pathologic signals considerably decreased and disturbed [3]. Accordingly, CWT and wavelet-like transforms provide more efficient features for diagnosis [4]. Analysis of pathological speech by CWT can also provide a visual pattern, which has a considerable help in diagnostics. The CWT-based analysis yields sharp and well-defined patterns which facilitate the diagnosis of speech disorders as reported in [5].

An extensive study is presented in [3] for identification of different voice disorders due to problems in the vocal folds. WPT is tested in [3] as features for their ability to analyze a signal at several levels of resolution. The performance of each WPT structure is evaluated in terms of the accuracy, sensitivity, specificity, and area under the receiver operating curve (AUC). Eventually, entropy features in the sixth level of WPT decomposition along with feature dimension reduction and an SVM-based classification method give the best results. WPT-based results do

M. H. Farouk, *Application of Wavelets in Speech Processing*, SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6_13,

better than previous works both in respect of accuracy and reduction of residues, which may lead to full accuracy and high-speed diagnosis procedure.

Another algorithm to discriminate voice disorders from each other is introduced in [6] using adaptive growth of WPT, based on the criterion of Local Discriminant Bases (LDB). A Mel-scaled WPT provides additional robust features of speech signals for automatic voice disorder diagnosis in [7] and [8]. Further set of features is based on WPT and singular value decomposition (SVD) for the detection of vocal fold pathology in [2]. The experimental results show very promising classification accuracy WPT-based features in [2].

For pitch detection, normal speech estimators are applied to pathological speech in [9]. A comparison is conducted among autocorrelation, harmonic product spectrum, and wavelet methods. Assessment of roughness, hoarseness, and breathiness was best with wavelets [1].

References

1. L. Baghai-Ravary, S.W. Beet, *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders, (SpringerBriefs in Electrical and Computer Engineering/ SpringerBriefs in Speech Technology)* (Springer, New York, 2013)
2. M. Hariharan, K. Polat, S. Yaacob, A new feature constituting approach to detection of vocal fold pathology. Int. J. Syst. Sci. http://www.tandfonline doi http://www.tandfonline.com/doi/full/10.1080/00207721.2013.794905#.UpmzASeAqWc. Accessed 14 May 2013
3. M. Arjmandi, M. Pooyan, An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. Biomed. Signal Process. Control **7**(1), 3–19 (2012)
4. P. Kukharchik, D. Martynov, I. Kheidorov, O. Kotov, Vocal fold pathology detection using modified wavelet-like features and support vector machines, in *15th European Signal Processing Conference (EUSIPCO 2007)* (2007), pp. 2214–2218
5. B. Ziółko, W. Kozłowski, M. Ziółko, R. Samborski, D. Sierra, J. Gałka, Hybrid wavelet-fourier-HMM speaker recognition. Int. J. Hybrid Inf. Technol. **4**(4), 25–42 (2011)
6. P.T. Hosseini, F. Almasganj, T. Emami, R. Behroozmand, S. Gharibzade, F. Torabinezhad, Local discriminant wavelet packet ‘ for voice pathology classification, in *The 2nd International Conference on Bioinformatics and Biomedical Engineering, 2008. ICBBE*, (2008), pp. 2052–2055
7. P. Murugesapandian, S. Yaacob, M. Hariharan, Feature extraction based on mel-scaled wavelet packet transform for the diagnosis of voice disorders, in *4th Kuala Lumpur International Conference on Biomedical Engineering 2008, BIOMED 2008*, vol. 21 (2008), pp. 790–793
8. M.P. Paulraj, Y. Sazali, M. Hariharan, Diagnosis of voice disorders using Mel scaled WPT and functional link neural network. Biomed. Soft Comput. Human Sci. **14**(2), 55–60 (2009)
9. C. Llerena, L. Alvarez, D. Ayllon, Pitch detection in pathological voices driven by three tailored classical pitch detection algorithms, in *Recent Advances in Signal Processing, Computational Geometry and System Theory: Proc ISCGAV’11 and ISTASC’11*, (2011), pp. 113–118

Index

M. H. Farouk, *Application of Wavelets in Speech Processing*,
SpringerBriefs in Speech Technology, DOI: 10.1007/978-3-319-02732-6,

Z