INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

A REVIEW ON LOW BIT RATE SPEECH CODING

RAVI KANT

Assistant Prof. Electrical Deptt.
IITM Murthal, Sonipat, Haryana

ABSTRACT: *As digital computers and communication systems continue to spread through our modern society, the use of digitized speech signals is increasingly common. The large number of bits required for the accurate reproduction of the speech waveform makes many of these systems complexes and expensive, so more efficient encoding of speech signals is desirable. For example, limited radio bandwidth is a major constraint in design of the next generation of public mobile telephone systems, and the speech data rate directly influences the bandwidth requirement. Also, computer storage of speech, such as in voice mail or voice response systems, becomes cheaper if the number of bits required for speech storage can be reduced. These are just some of the applications which can benefit from the development of algorithms to significantly reduce the speech data rate. In this paper a review of low bit rate speech coding is given.*

Keywords: *Speech Coding, Bit Rat, Vocoders, Very Low Bit Rate.*

1. INTRODUCTION

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission over band-limited wired and wireless channels and/or storage [2]. Today, speech coders have become essential components in telecommunications and in the multimedia infrastructure. Commercial systems that rely on efficient speech coding include cellular communication, voice over internet protocol (VOIP), videoconferencing, electronic toys, archiving, and digital simultaneous voice and data (DSVD), as well as numerous PC-based games and multimedia applications. Speech coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stored in digital media, and decoding the signal with the best possible perceptual quality. Like any other continuous-time signal, speech may be represented digitally through the processes of sampling and quantization; speech is typically quantized using either 16-bit uniform or 8-bit companded quantization. Like many other signals, however, a sampled speech signal contains a great deal of information that is either redundant (nonzero mutual information between successive samples in the signal) or perceptually irrelevant (information that is not perceived by human listeners). Most telecommunications coders are *lossy*, meaning that the synthesized speech is perceptually similar to the original but may be physically dissimilar [2]. The costs of digital storage and transmission media are generally proportional to the amount of digital data that can be stored or transmitted. While the cost of such media decreases every year, the demand for their use increases at an even higher rate. Therefore, there is a continuing need to minimize the number of bits necessary to transmit signals while maintaining acceptable signal fidelity or quality [9].

Low-bit-rate speech coding, at rates below 4 kb/s, is needed for both communication and voice storage applications. At such low rates, full encoding of the speech waveform is not possible; therefore, low-rate coders rely instead on parametric models to represent only the most perceptually-relevant aspects of speech [1]. While there are a number of different approaches for this modeling, all can be related to the basic linear model of speech production, where an excitation signal drives a vocal tract filter. The basic properties of the speech signal and of human speech perception can explain the principles of parametric speech coding as applied in early vocoders. Current speech modeling approaches, such as mixed excitation linear prediction, sinusoidal coding, and waveform interpolation, use more sophisticated versions of these same concepts. Modern techniques for encoding the model parameters, in particular using the theory of vector quantization, allow the encoding of the model information with very few bits per speech frame. Successful standardization of low-rate coders has enabled their widespread use for both military and satellite communications, at rates from 4 kb/s all the ways down to 600 b/s. However, the goal of toll quality low-rate coding continues to provide a research challenge [1]. In order to achieve bit rates lower than 600 bps in speech coding, it is necessary to use recognition and synthesis techniques. By transmitting only the indexes of the recognized unit, the transmission bit rate is drastically reduced [6]. For very low data rates, realistic experiments have shown that vector quantization can achieve a given level of average distortion with 15 to 20 fewer bits/frame than that required for the optimized scalar quantizing approaches presently in use [7]. With low bit-rate coders, MSE does not give a valid estimate of the perceptual errors and perceptive distortion measures might be more useful [10].

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

DIGITAL REPRESENTATION OF SPEECH SIGNALS:

In general speech coding may be defined as a process that generates sequences of binary digits from the speech signal. The goal of the modern-day speech coders is to devise a compact digital representation of the speech signal that will enable their transmission through band-limited channels and yield a perceptually acceptable reconstruction at the receiver [8]. Speech signal sampled at 8000 Hz and each sample quantized to 8 bits results in a data-rate of 64 Kbps. The quality of the speech thus represented is is indistinguishable from the 4 KHz band-limited analog speech and often referred to as *broadcast* quality speech [8]. This sampled and quantized representation of the speech is typically obtained at the output of analog to digital converters and forms the primary signal input to all speech processing algorithms. The sampled and quantized speech signal (henceforth just referred to as the original speech signal) is seldom transmitted directly at 64 Kbps. A significant reduction in transmitted data-rate at the cost of a marginal degradation in the perceptual quality of the reconstructed speech can be achieved by employing a speech coder. Speech coders may broadly be classified into three groups: *waveform coders*, *parametric coders* (or model based) and *hybrid coders* [8]. Waveform coders seek to represent the waveform of the speech signal using digital symbols and typically operate at bit-rates in the 16{64 Kbps range. To achieve speech coding at bit-rates below 16 Kbps, a source {system model [18] for the generation of the speech signals is often assumed. Such a model seeks to capture the perceptually significant information in a speech signal by modeling it as the output of an autoregressive system whose input is an excitation signal inspired by the mechanism of generation of speech by humans [8]. Fully parametric model based coders compress the speech signal by efficiently encoding the parameters of the autoregressive model and the source for speech generation. At the decoder, the received parameters are used to reconstruct the speech generation model, which is then used to synthesize the speech signal. The efficiency of such a speech coding system largely depends on the success of the model in representing the perceptually important components of the speech signal. While fully parametric coders can achieve speech compression to bit-rates below 8 Kbps, the quality of the synthesized speech is often poor due to deficiencies in the models [8]. The hybrid coders seek to balance the trade between the bit-rate and the speech reconstruction quality. While hybrid coders work within the paradigm of the source filter model, they still try to match the speech signal waveform to the output of the speech model. Typically this is done by performing analysis of the speech signal to estimate the model parameters via synthesis. Such coders encode speech at bit-rates in the 8{16 Kbps and achieve a quality similar to that of waveform coders [8].

2. HUMAN SPEECH PERCEPTION

Human perception of speech is determined by the capabilities of the human auditory system, which consists of the ear, the auditory nerve, and the brain [1]. The ear serves as a transducer, converting the acoustic input at the outer ear first to bone vibrations in the middle ear, then to fluid motion in the cochlea of the inner ear, and finally to electrical pulses generated by the inner hair cells in the cochlea. The location of maximum fluid vibration in the cochlea varies systematically with the input signal frequency, and this frequency response varies with the strength of the input signal. Also, the inner hair cells only detect motion in one direction. Thus, the ear acts like a very large bank of band-pass filters with dynamic range compression, and the output of each filter undergoes half-wave rectification. The half-wave rectified band-pass filter outputs are transmitted across the auditory nerve to the lower levels of the brain, where specialized neurons can perform basic signal processing operations. For example, there are neurons that respond to onsets and to decays, and other neurons may be able to estimate autocorrelations by comparing a signal to a delayed version of itself. The outputs of these neurons are then passed to higher levels of the brain for more sophisticated processing. This results in the final analysis of the acoustic signal based on context and additional knowledge, such as classification of sound source and interpretation of pattern. In speech processing, this includes recognition of words as well as analysis of speaker identity and emotional state [1]

3. SPEECH PRODUCTION

Speech is a sequence of sounds generated by the human vocal system [1]. The acoustic energy necessary to produce speech is generated by exhaling air from the lungs. This air stream is used to produce sound in two different ways: by vibrating the vocal cords or by forcing air turbulence. If the vocal cords are used, the speech is referred to as voiced speech; otherwise, the speech is called unvoiced. In voiced speech, the opening and closing of the vocal cords at the glottis produces quasi-periodic puffs of air called glottal pulses, which excite the acoustic tubes of the vocal and nasal tracts. The average spacing between glottal pulses is called the pitch period. The frequency content of the

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

resulting acoustic wave propagating from the mouth depends on the spectrum of the glottal pulses and on the configuration of the vocal tract [1].

4. SPEECH CODER ATTRIBUTES

Speech quality as produced by a speech coder is a function of bit rate, complexity, delay, and bandwidth. Hence, when considering speech coders it is important to review all these attributes [3]. It is important to realize that there is a strong interaction between all these attributes and that they can be traded off against each other. For example, low-bit-rate coders tend to have more delay than higher-bit-rate coders. They may also require higher complexity to implement and often have lower quality than the higher-bit-rate coders. In the remainder of this article we limit ourselves to telephone bandwidth speech (200-3400 Hz) sampled at 8kWz. Additional factors that influence the selection of a given speech coder are availability and licensing conditions, or it could be the way the standard is specified. Some standards are only described as an algorithmic description, while others are defined by bit exact American National Standards Institute ANSI-C code [3].

5. BIT RATE

Bit rate is the simplest attribute to understand, but is less straight forward than one might imagine. Most speech coders operate at a fixed bit rate regardless of the input signal characteristics. Since multimedia speech coders share the channel with other forms of data, it is better to make the coder variable- rate. For simultaneous voice and data applications, a good compromise is to create a silence compression scheme as part of the coding standard. A common solution is to use a fixed rate for active speech and a low rate for background noise [3]

6. PRINCIPLES OF VERY LOW BIT RATE (VLBR) SPEECH CODING TRAINING PHASE

An unsupervised training phase is used to build the HMM models and the codebook of synthesis units. During the initial step, spectral target vectors and corresponding segmentation are obtained through Temporal Decomposition (TD) of the training speech corpus. Vector Quantisation (VQ) is then used to cluster the different segments in a limited number of classes (64). Finally, for each class of segments, 3-states left-to-right HMM (Hidden Markov Model) models are trained using an iterative process refining both the segmentation and the estimation of the HMM models. The final segmentation is obtained with the final set of HMM models, and is used to build the reference codebook of synthesis units. More details on the training process can be found in [4].

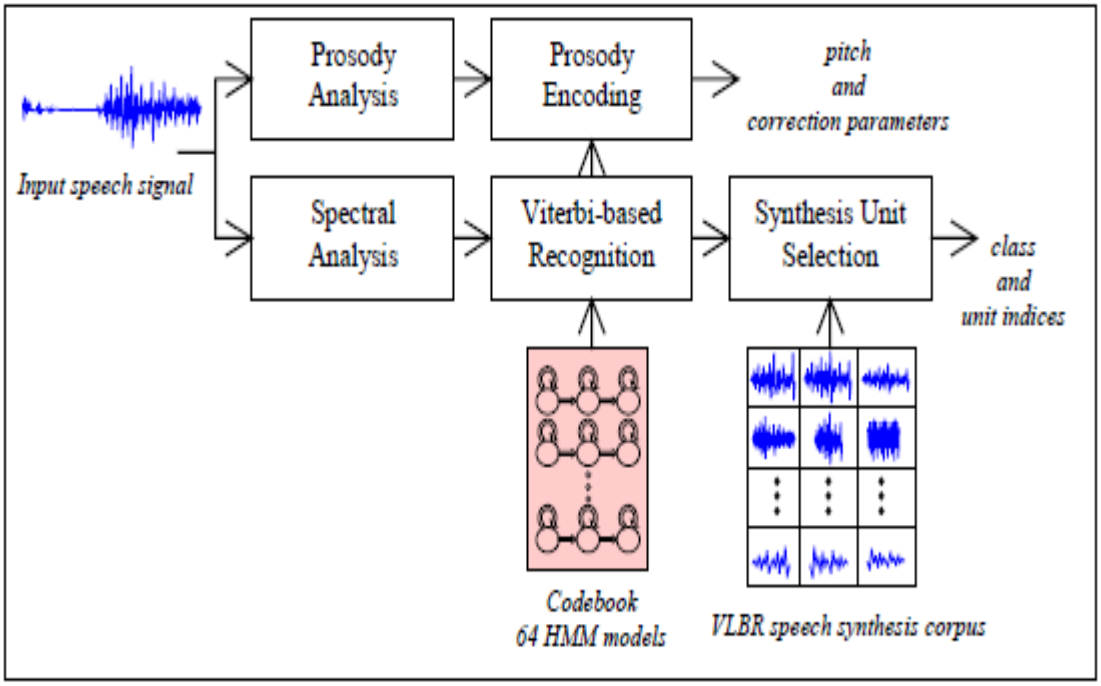

Figure 1: VLBR encoding principle [4]

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

7. ENCODING PHASE

During the encoding phase, a Viterbi algorithm provides the on-line segmentation of speech using the previously trained HMM models, together with the corresponding labeling as a sequence of class (or HMM) indices. Each segment is then further analyzed in terms of prosody profile: frame-based evolution of pitch and energy values. The unit selection process is finally used to find an optimal synthesis unit in the reference codebook. In order to take into account the backward context information, each class of the synthesis codebook is further organized in sub-classes, depending on the previous identified class [4].

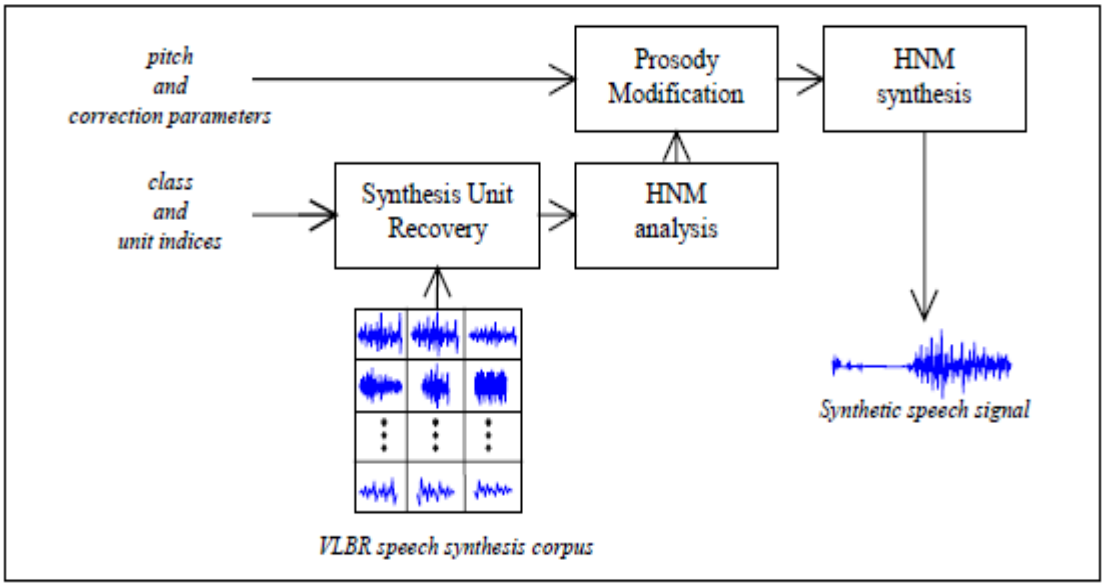

Figure 2: VLBR decoding principle [4]

8. DECODING PHASE

During the decoding phase, the synthesis units are recovered from the class and unit indices and concatenated with a HNM-like algorithm (Harmonic plus Noise Model). Additional parameters characterising the prosody information are also incorporated to match the original speech signal.

9. PERFORMANCE

In digital communications, speech quality is classified into four general categories, namely: broadcast, network or toll, communications, and synthetic. Broadcast wideband speech refers to high-quality "commentary" speech that can generally be achieved at rates above 64k bits ls. Toll or network quality refers to quality comparable to the classical analog speech (200-3200 Hz) and can be achieved at rates above 16 kbits/s. Communications quality implies somewhat degraded speech quality which is nevertheless natural, highly intelligible, and adequate for telecommunications. Synthetic speech is usually intelligible but can be unnatural and associated with a loss of speaker recognizability. Communications speech can be achieved at rates above 4.8 kbitsls and the current goal in speech coding is to achieve communications quality at 4.0k bits ls. Currently, speech coders operating well below 4.0k bits ls tend to produce speech of synthetic quality [5].

10. CONCLUSION & FUTURE SCOPE

Speech coding is the process of obtaining a compact representation of voice signals for proficient transmission over band-limited wired and wireless channels and/or storage. Low-bit-rate speech coding, at rates below 4 kb/s, is needed for both communication and voice storage applications. At such low rates, full encoding of the speech waveform is not possible; therefore, low-rate coders depend instead on parametric models to represent only the most perceptually-relevant aspects of speech. The basic properties of the speech signal and of human speech perception can explain the principles of parametric speech coding as applied in early vocoders. Current speech modeling approaches, such as mixed excitation linear prediction, sinusoidal coding, and waveform interpolation, use more sophisticated versions of these same concepts. Modern

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

techniques for encoding the model parameters, in particular using the theory of vector quantization, allow the encoding of the model information with very few bits per speech frame. In future we can further reduce the bit rate.

REFERENCES

[1] Alan McCree "Low-Bit-Rate Speech Coding" Springer Handbook on Speech Processing and Speech Communication. Pp. 1-30.

[2] Mark Hasegawa-Johnson and Abeer Alwan "Speech Coding: Fundamentals and Applications" Wiley Encyclopedia of Telecommunications, Edited by John G. Proakis ISBN 0-471-36972-1. Pp. 1-20.

[3] Richard V.Cox and Peter Karoon"Low Bit Rate Speech Coders for Multimedia Communication" IEEE Communicatiuns Magazine 0 December 1996.pp. 34-41.

[4] Marc Padellini, François Capman, and Geneviève Baudoin "Very Low Bit Rate (Vlbr) Speech Coding Around 500 Bits/Sec". Thales Communications, 160, Bd de Valmy , BP 82, 92704 Colombes, CEDEX.pp. 1669-1672.

[5] Andreas S. Spanias "Speech Coding: A Thtorial Review" Proceedings of the IEEE, Vol. 82. No. 10, October 1994. Pp. 1541-1584.

[6] Geneviève Baudoin, Françoi Capman, Jan Cˇ ernocký, Fadi El Chami, Maurice Charbit4, Gérard Chollet, and Dijana Petrovska-Delacrétaz "Advances in Very Low Bit Rate Speech Coding Using Recognition and Synthesis Techniques" LNAI 2448, pp. 269–276, 2002.

[7] Andres Buzo, Augustine H. *Gray*, Jr., Robert M. Gray, , and John D. Markel "Speech Coding Based Upon Vector Quantization" IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-28, No. *5*, October 19813. Pp. 562-574.

[8] Venkatesh Krishnan "A Framework For Low Bit-Rate Speech Coding In Noisy Environment" School of Electrical and Computer Engineering Georgia Institute of Technology March 2005.

[9] John Makhoul, Salim Roucos and Herbert Gish "Vector Quantization in Speech Coding" Proceedings of the IEEE, Vol. 73. NO 11. **November 1985. Pp. 1550-1588.**

[10] Bruno Carpentieri "Low Bit Rate Speech Coding via TCVRQ" Recent Advances in Computer Engineering, Communications and Information Technology. ISBN: 978-960-474-361-2. Pp. 126-131.