Prolog to

Speech Coding: A Tutorial Review

A tutorial introduction to the paper by Spanias

If the subtle sounds of human speech are to travel the information highways of the future, digitized speech will have to be more efficiently transmitted and stored. Designers of cellular communications systems, wireless personal computer networks, and multimedia systems are all searching for improved techniques for handling speech.

Since its awkward beginnings in the 1930's, speech coding has developed to become an essential feature of everyday telephone system operations. Speech coding is now finding applications in cellular communications, computer systems, automation, military communications, biomedical systems, and almost everywhere that digital communication takes hold.

Speech coding involves sampling and amplitude quantization of the speech signal. The aim is to use a minimum number of bits, while preserving the quality of the reconstructed speech at the receiving end. Coding research is now taking aim at low-rate (8 to 2.4 kbits/s) and very-low-rate (below 2.4 kbits/s) techniques.

The entire gamut of speech coding research is covered in this paper. An extensive list of references gives the reader access to the speech coding literature. The paper has tutorial information to orient applications engineers, and it nicely summarizes coder developments for research experts.

The meaning of the words we speak often changes with the smallest inflection of our voices, so better speech quality is an essential goal for coding research. The paper lays out the quality levels of reconstructed speech, ranging from the highest quality broadcast, wide-band speech produced by coders at 64 kbits/s, to the lowest quality, synthetic speech, currently produced by coders that operate well below 4 kbits/s. A section on speech quality points out that subjective testing can be lengthy and costly. Speech quality has been gauged by objective measures, beginning with the signal-to-noise ratio, but these measures do not account for human perception.

The bulk of this paper is devoted to explaining and reviewing a wide variety of speech coders. First of all, the paper discusses waveform coders. Waveform coders, as opposed to vocoders, compress speech waveforms without making use of the underlying speech models.

Scalar quantization techniques include familiar classical methods such as pulse-code modulation (PCM), differential PCM, and delta modulation.

Vector quantization techniques make use of codebooks that reside in both the transmitter and receiver. The paper attributes much of the progress recently achieved in low-rate speech coding to the introduction of vector quantization techniques in linear predictive coding. Highly structured codebooks allow significant reduction in the complexity of high-dimensional vector quantization.

Sub-band and transform coders rely on transform-domain representations of the voice signal. In sub-band coders, these representations are obtained through filter banks. Sub-band encoding is used in medium-rate coding. Fourier transform coders obtain frequency-domain representations by using unitary transforms. Perhaps the most successful of the early transform coders is the adaptive transform coder was developed at Bell Laboratories.

The paper describes analysis–synthesis methods that use the short-time Fourier transform, and also various methods that use sinusoidal representations of speech. Multiple sine waves have been successfully used in many different speech coding systems. For example, one sinusoidal analysis–synthesis system performed very well with a variety of signals, including those from multiple speakers, music, and biological sounds, and this system also performed well in the presence of background noise. Sinusoidal coders have been used for low-rate speech coding, and have produced high-quality speech in the presence of background noise. Another coder that belongs to this class is the multiband excitation coder which recently became part of the Australian mobile satellite and International Mobile Satellite standards.

Since 1939, vocoder systems have tried to produce intelligible human speech without necessarily matching the speech waveform. Initially, simple models were used to produce low-rate coding. The result was synthetic, buzzy-sounding reconstructed speech. More recently, sophisticated vocoders have provided improved quality at the cost of increased complexity. The paper briefly describes channel and formant vocoders, and the homomorphic vocoder, but focuses mostly on linear predictive vocoders.

0018–9219/94$04.00 © 1994 IEEE

Linear predictive coders use algorithms to predict the present speech sample from past samples of speech. Usually 8 to 14 linear predictive parameters are required to model the human vocal tract. The analysis window is typically 20–30 ms long and parameters are generally updated every 10–30 ms. Real-time predictive coders were first demonstrated in the early 1970's. The paper describes a linear predictive coding algorithm that has become a U.S. federal standard for secure communications at 2.4 kbits/s. The U.S. Government is currently seeking an improved algorithm to replace that standard.

In analysis-by-synthesis methods, the reconstructed and original speech are compared, and the excitation parameters are adjusted to minimize the difference before the code is transmitted.

Hybrid coders determine speech spectral parameters by linear prediction and optimize excitation using analysis-by-synthesis techniques. These hybrid coders combine the features of modern vocoders with an ability to exploit the properties of the human auditory system. The paper describes several analysis-by-synthesis linear predictive coding algorithms. The coder used in the British Telecom International skyphone satellite-based system is based on one of these algorithms (MPLP). Another of these algorithms (RPE-LTP) has been adopted for the full-rate GSM Pan-European digital mobile standard. The U.S. Department of Defense has adopted another algorithm (CELP) described in the paper, for possible use in a new secure telephone unit. The 8-kbits/s algorithm (VSELP) adopted for the North American Cellular Digital System is also described, as is the LD-CELP coder selected by the CCITT as its recommendation for low-delay speech coding.

—Howard Falk

Speech Coding: A Tutorial Review

ANDREAS S. SPANIAS, MEMBER, IEEE

The past decade has witnessed substantial progress towards the application of low-rate speech coders to civilian and military communications as well as computer-related voice applications. Central to this progress has been the development of new speech coders capable of producing high-quality speech at low data rates. Most of these coders incorporate mechanisms to: represent the spectral properties of speech, provide for speech waveform matching, and "optimize" the coder's performance for the human ear. A number of these coders have already been adopted in national and international cellular telephony standards.

The objective of this paper is to provide a tutorial overview of speech coding methodologies with emphasis on those algorithms that are part of the recent low-rate standards for cellular communications. Although the emphasis is on the new low-rate coders, we attempt to provide a comprehensive survey by covering some of the traditional methodologies as well. We feel that this approach will not only point out key references but will also provide valuable background to the beginner. The paper starts with a historical perspective and continues with a brief discussion on the speech properties and performance measures. We then proceed with descriptions of waveform coders, sinusoidal transform coders, linear predictive vocoders, and analysis-by-synthesis linear predictive coders. Finally, we present concluding remarks followed by a discussion of opportunities for future research.

I. INTRODUCTION

Although with the emergence of optical fibers bandwidth in wired communications has become inexpensive, there is a growing need for bandwidth conservation and enhanced privacy in wireless cellular and satellite communications. In particular, cellular communications have been enjoying a tremendous worldwide growth and there is a great deal of R&D activity geared towards establishing global portable communications through wireless personal communication networks (PCN's). On the other hand, there is a trend toward integrating voice-related applications (e.g., voice-mail) on desktop and portable personal computers—often in the context of multimedia communications. Most of these applications require that the speech signal is in digital format so that it can be processed, stored, or transmitted under software control. Although digital speech brings flexibility and opportunities for encryption, it is also associated (when uncompressed) with a high data rate and hence high requirements of transmission bandwidth and storage. *Speech Coding* or *Speech Compression* is the field concerned with obtaining compact digital representations of voice signals for the purpose of efficient transmission or storage. Speech coding involves sampling and amplitude quantization. While the sampling is almost invariably done at a rate equal to or greater than twice the bandwidth of analog speech, there has been a great deal of variability among the proposed methods in the representation of the sampled waveform. The objective in speech coding is to represent speech with a minimum number of bits while maintaining its perceptual quality. The quantization or binary representation can be direct or parametric. Direct quantization implies binary representation of the speech samples themselves while parametric quantization involves binary representation of speech model and/or spectral parameters.

With very few exceptions, the coding methods discussed in this paper are those intended for digital speech communications. In this application, speech is generally bandlimited to 4 kHz (or 3.2 kHz) and sampled at 8 kHz. The simplest nonparametric coding technique is Pulse-Code Modulation (PCM) which is simply a quantizer of sampled amplitudes. Speech coded at 64 kbits/s using logarithmic PCM is considered as "noncompressed" and is often used as a reference for comparisons. In this paper, we shall use the term *medium rate* for coding in the range of 8–16 kbits/s, *low rate* for systems working below 8 kbits/s and down to 2.4 kbits/s, and *very low rate* for coders operating below 2.4 kbits/s.

Speech coding at medium-rates and below is achieved using an *analysis–synthesis* process. In the analysis stage, speech is represented by a compact set of parameters which are encoded efficiently. In the synthesis stage, these parameters are decoded and used in conjunction with a reconstruction mechanism to form speech. Analysis can be *open-loop* or *closed-loop*. In closed-loop analysis, the parameters are extracted and encoded by minimizing explicitly a measure (usually the mean square) of the difference between the original and the reconstructed speech. Therefore, closed-loop analysis incorporates synthesis and hence this process is also called *analysis by synthesis*. Parametric representations can be speech- or non-speech-specific. Non-speech-specific coders or waveform coders are concerned with the

Manuscript received July 6, 1993; revised March 4, 1994. Portions of this work have been supported by Intel Corporation.

The author is with the Department of Electrical Engineering, Telecommunications Research Center, Arizona State University, Tempe, AZ 85287-5706 USA.

IEEE Log Number 9401511.

faithful reconstruction of the time-domain waveform and generally operate at medium rates. Speech-specific coders or voice coders (*vocoders*) rely on speech models and are focussed upon producing perceptually intelligible speech without necessarily matching the waveform. Vocoders are capable of operating at very-low rates but also tend to produce speech of synthetic quality. Although this is the generally accepted classification in speech coding, there are coders that combine features from both categories. For example, there are speech-specific waveform coders such as the Adaptive Transform Coder [303] and also hybrid coders which rely on analysis-by-synthesis linear prediction. Hybrid coders combine the coding efficiency of vocoders with the high-quality potential of waveform coders by modeling the spectral properties of speech (much like vocoders) and exploiting the perceptual properties of the ear, while at the same time providing for waveform matching (much like waveform coders). Modern hybrid coders can achieve communications quality speech at 8 kbits/s and below at the expense of increased complexity. At this time there are at least four such coders that have been adopted in telephony standards.

A. Scope and Organization

In this paper, we provide a survey of the different methodologies for speech coding with emphasis on those methods and algorithms that are part of recent communications standards. The paper is intended both as a survey and a tutorial and has been motivated by advances in speech coding which have enabled the standardization of low-rate coding algorithms for civilian cellular communications. The standardizations are results of more than fifty years of speech coding research. Until recently, low-rate algorithms were of interest only to researchers in the field. Speech coding is now of interest to many engineers who are confronted with the difficult task of learning the essentials of voice compression in order to solve implementation problems, such as fitting an algorithm to an existing fixed-point signal processor or developing low-power single-chip solutions for portable cellular telephones. Modern speech-coding algorithms are associated with numerical methods that are computationally intensive and often sensitive to machine precision. In addition, these algorithms employ mathematical, statistical, and heuristic methodologies. While the mathematical and statistical techniques are associated with the theory of signal processing, communications, and information theory, many of the heuristic methods were established through years of experimental work. Therefore, the beginner not only has to get a grasp of the theory but also needs to review the algorithms that preceded the standards. In this paper we attempt to sort through the literature and highlight the key theoretical and heuristic techniques employed in classical and modern speech-coding algorithms. For each method we give the key references and, when possible, we refer first to the article that the novice will find more accessible.

The general notation adopted in this paper is as follows. The discrete-time speech signal is denoted as $s(n)$, where n is an integer indexing the sample number. Discrete-time speech is related to analog speech, $s_a(t)$, by $s(n) = s_a(nT) = s_a(t)|_{t=nT}$, where T is the sampling period. Unless otherwise stated, lower case symbols denote time-domain signals and upper case symbols denote transform-domain signals. Bold characters are used for matrices and vectors. The rest of the notation is introduced in subsequent sections as necessary.

The organization of the paper is as follows. The first section gives a brief description of the properties of speech signals and continues with a historical perspective and a review of performance measures. In Section II, we discuss waveform coding methods. In particular, we start with a general description of scalar [55], [82], [152] and vector quantization [81], [98], [115], [192] methods and we continue with a discussion of waveform coders [48], [52]. Section III presents sinusoidal analysis–synthesis methods [205] for voice compression and Section IV presents vocoder methods [11], [162], [163]. Finally, in Section V we discuss analysis-by-synthesis linear predictive coders [96], [100], [123], [272] and in Section VI we present concluding remarks. Low-rate coders, and particularly those adopted in the recent standards, are discussed in more detail. The scope of the paper is wide and although our literature review is thorough is by no means exhaustive. Papers with similar scope [12], [23], [82], [83], [96], [104], [109], [150], [154], [155], [157], [191], [270], [279]; special journal and magazine editions on voice coding [18], [19], [131], [132], [134]–[136], [138], [139]; and books on speech processing [62], [86], [90], [91], [99], [113], [152], [199], [232], [234], [236], [251], [275] can provide additional information. There are also six excellent collections of papers edited by Jayant [156], Davidson and Gray [61], Schafer and Markel [269], Abut [1], and Atal, Cuperman, and Gersho [9], [10]. For the reader, who wants to keep up with the developments in this field, articles appear frequently in IEEE TRANSACTIONS and symposia associated with the areas of signal processing and communications (see references section) and also in specialized conferences, workshops, and journals, e.g., [133] [137], [140], [291].

B. Speech Properties

Before we begin our presentation of the speech coding methods, it would be useful if we briefly discussed some of the important speech properties. First, speech signals are nonstationary and at best they can be considered as quasi-stationary over short segments, typically 5–20 ms. The statistical and spectral properties of speech are thus defined over short segments. Speech can generally be classified as voiced (e.g., /a/, /i/, etc), unvoiced (e.g., /sh/), or mixed. Time- and frequency-domain plots for sample voiced and unvoiced segments are shown in Fig. 1. Voiced speech is quasi-periodic in the time domain and harmonically structured in the frequency domain while unvoiced speech is random-like and broadband. In addition, the energy of voiced segments is generally higher than the energy of unvoiced segments.

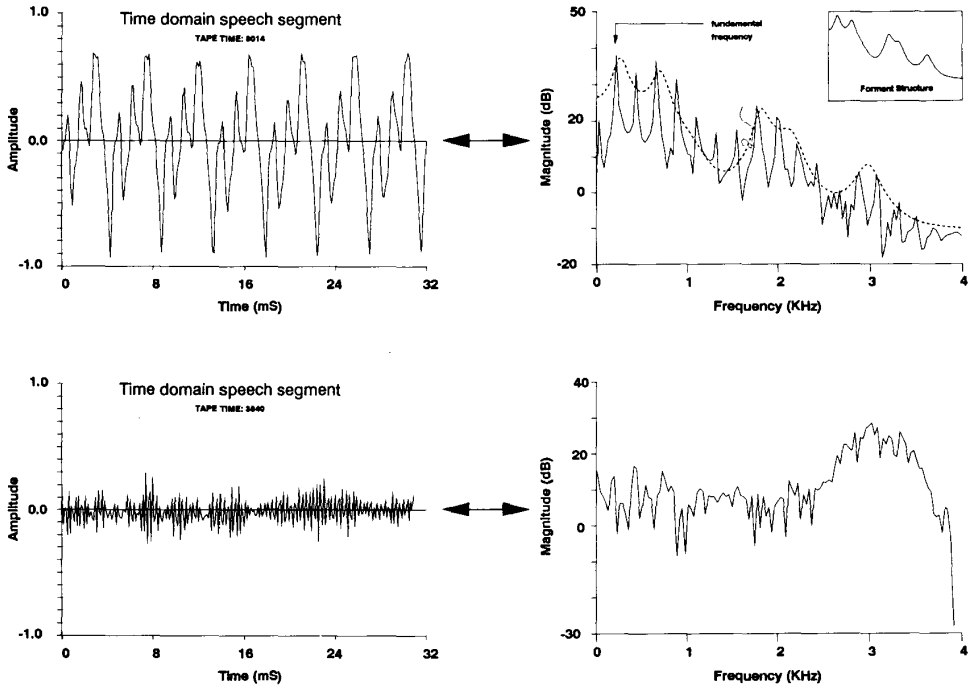

Fig. 1. Voiced and unvoiced segments and their short-time spectra.

The short-time spectrum[1] of voiced speech is characterized by its *fine* and *formant* structure. The fine harmonic structure is a consequence of the quasi-periodicity of speech and may be attributed to the vibrating vocal chords. The formant structure (spectral envelope) is due to the interaction of the source and the vocal tract. The vocal tract consists of the pharynx and the mouth cavity. The shape of the spectral envelope that "fits" the short-time spectrum of voiced speech, Fig. 1, is associated with the transfer characteristics of the vocal tract and the spectral tilt (6 dB/octave) due to the glottal pulse [261]. The spectral envelope is characterized by a set of peaks which are called formants. The formants are the resonant modes of the vocal tract. For the average vocal tract there are three to five formants below 5 kHz. The amplitudes and locations of the first three formants, usually occurring below 3 kHz, are quite important both in speech synthesis and perception. Higher formants are also important for wideband and unvoiced speech representations. The properties of speech are related to the physical speech production system as follows. Voiced speech is produced by exciting the vocal tract with quasi-periodic glottal air pulses generated by the vibrating vocal chords. The frequency of the periodic pulses is referred to as the fundamental frequency or pitch. Unvoiced speech is produced by forcing air through a constriction in the vocal tract. Nasal sounds (e.g., /n/) are due to the acoustical coupling of the nasal tract to the vocal tract, and plosive sounds (e.g., /p/) are produced by abruptly releasing air pressure which was built up behind a closure in the tract.

[1] Unless otherwise stated the term spectrum implies power spectrum

More information on the acoustic theory of speech production is given by Fant [75] while information on the physical modeling of the speech production process is given in the classic book by Flanagan [86].

C. Historical Perspective

Speech coding research started over fifty years ago with the pioneering work of Homer Dudley [66], [67] of the Bell Telephone Laboratories. The motivation for speech coding research at that time was to develop systems for transmission of speech over low-bandwidth telegraph cables. Dudley practically demonstrated the redundancy in the speech signal and provided the first analysis–synthesis method for speech coding. The basic idea behind Dudley's voice coder or vocoder (Fig. 2) was to analyze speech in terms of its pitch and spectrum and synthesize it by exciting a bank of ten analog band-pass filters (representing the vocal tract) with periodic (buzz) or random (hiss) excitation (for voiced and unvoiced sounds, respectively). The channel vocoder received a great deal of attention during World War II because of its potential for efficient transmission of encrypted speech. Formant [223] and pattern matching [68] vocoders along with improved analog implementations of channel vocoders [221], [292] were reported through the 1950's and 1960's. In the formant vocoder, the resonant characteristics of the filter bank track the movements of the formants. In the pattern-matching vocoder the best match between the short-time spectrum of speech and a set of stored frequency response patterns is determined and speech is produced by exciting the channel filter associated with the selected pattern. The pattern-matching vocoder was

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.