SPEECH SIGNAL PROCESSING

Bastiaan Kleijn
Professor of
Speech Signal Processing

During the year 2001, the Speech Signal Processing Group at the Department of Speech, Music and Hearing at KTH consisted of seven Ph.D. students (five of whom were located at the department), a part-time (20%) researcher (forskarassistent), two guest researchers, and a professor. The group performs research encompassed within speech processing, signal processing, and source coding and teaches two undergraduate courses (Information Theory and Source Coding, and Digital Speech Signal Processing), in addition to a varying number of graduate courses. The group also supervises numerous 5-month projects performed by undergraduate students.

The research of the group is mostly aimed towards improved algorithms for speech and audio coding, speech synthesis, and speech enhancement for various applications. In general, research in these areas has made great strides in the last few decades and the results of this labor are now part of everyday life. Speech coding is an enabling technology for mobile telephones. Audio coding is becoming commonplace in consumer electronic devices. Speech synthesis is often used in tele-communication services and speech enhancement is used for communications in adverse environments. Despite these recent advances, increasing quality and lowering bit rate remain important challenges for the future. New communication network technologies have introduced fresh challenges such as wideband speech coding, voice and audio over the Internet

to name only two. In the following, we provide a brief overview of the main research activities of the group during the year 2001.

Speech coding and synthesis

In speech coding and synthesis, the work of the group aimed at two topics: i) the waveform interpolation (WI) algorithm, and ii) a new linear prediction vector quantization method for speech signals.

With respect to waveform interpolation, the group continued its research on improving the basic paradigm; this work is also relevant for sinusoidal coding. Conventional sinusoidal and waveform interpolation coders have a modeling error that limits performance at high rates. Their time-frequency localization of the unvoiced speech component is often insufficient to characterize the speech signal in a perceptually accurate manner with few components. We address these problems by using two frame expansions: one for the signal waveform, and a second one that describes the time evolution of the coefficients of the first one. The second frame expansion can be used to perform a voiced-unvoiced decomposition of the speech signal. The quality of such decomposition is affected by pitch fluctuations. We showed that using a novel continuous-time spline-based pitch estimation (optimization) algorithm the energy concentration associated with the second frame expansion can be vastly improved, especially near onsets. This improvement avoids errors in the voiced-unvoiced decompositions. Since errors in the voiced-unvoiced decomposition lead to a rapid increase of the bit rate, the scheme will lead to a lower average bit rate at a given quality level.

Our new linear prediction vector quantization method was motivated by the fact that, compared to a scalar quantizer (SQ), a vector quantizer (VQ) has memory, space filling, and shape advantages. If the signal statistics are known, direct vector quantization (DVQ) according to these statistics provides the highest coding efficiency, but requires unmanageable storage requirements if the statistics are time varying. In code-excited linear predictive (CELP) coding, a single "compromise" codebook is trained in the excitation-domain and the space filling and shape advantages of VQ are utilized in a non-optimal, average sense. We have proposed a new method where the space-filling, the memory, and shape advantages are all fully utilized. Our experiments show that the new method provides a higher SNR than CELP and (single-codebook) DVQ, and has a computational complexity similar to DVQ and much lower than CELP. Storage requirements are modest.

Audio coding

Traditionally, audio coders have used nonparametric descriptions of the signal based on filter banks. However, within the last five years parametric coding techniques have been shown to facilitate efficient coding of audio signals, particularly at low rates. We contribute in this area in a collaborative project with Delft University of Technology and Philips Research in Eindhoven, both in the Netherlands. We highlight specific areas that involved KTH work.

The joint project is arranged around the development of an efficient low-rate audio coder, which describes the audio signal using a set of signal models. Most important of the signal models is the sinusoidal model, which operates on a segmental basis. The sinusoidal model includes damped sinusoids. We continued our work on increasing the modeling efficiency of damped sinusoids by modifying the locations of signal transients. As a result of the modifications, transients only occur at the beginnings of sinusoidal segments. The modified signal is perceptually indistinguishable from the original signal. The modeling efficiency and reconstruction quality are increased significantly if the signal is modified as described above prior to sinusoidal modeling.

An important aspect of an audio coder is quantization of a signal representation. In sinusoidal modeling, the signal representation includes amplitude and phase of sinusoidal components. We have derived analytical formulas for amplitude and phase quantization using high-rate assumptions. Entropy-constrained quantization, which is relevant for statistical networks such as the Internet, is considered. We developed an entropy-constrained unrestricted polar quantization (UPQ) method, where phase quantization depends on the input amplitude. The UPQ is also generalized to include a weighted error measure, such that it accounts for masking effects of the human auditory system. In our method, both amplitude and phase quantization

depend on the perceptual importance of sinusoids. When applied to a sinusoidal audio coder, the new method outperforms conventional sinusoidal quantization where phase is quantized with a constant number of bits for all audible sinusoids.

Speech enhancement

Our work on speech enhancement included two topics: i) bandwidth extension of telephone bandwidth speech and ii) the estimation of speech model-parameters under environmental noise.

Telephone speech is usually limited to less than 4 kHz in bandwidth creating the typical sound of telephone speech. While it is well known that wide-band speech sounds significantly better than this narrow-band signal, the existing infrastructure has prevented the widespread introduction of wide-band signals. Thus, there is a strong motivation for bandwidth extension, i.e., the creation of wide-band speech from a narrow-band speech. In general the bandwidth extension schemes aim at finding a filter representing the wide-band signal as well as the excitation of this filter given only the narrow-band signal. The wide-band spectral envelope estimation is usually performed using a statistical mapping between narrow-band features and wide-band/high-band spectral envelope. The performance of the statistical mapping depends strongly on the shared information between the narrow- and high-band. We have investigated the dependency between the spectral envelopes of speech in disjoint frequency bands, one covering the telephone bandwidth from 0.3 kHz to 3.4 kHz and one covering the frequencies from 3.7 kHz to 8 kHz. The spectral envelopes were jointly modeled with a Gaussian mixture model based on mel-frequency cepstral coefficients and the log-energy-ratio of the disjoint frequency bands. Using this model, we have quantified the dependency between bands through their mutual information and the perceived entropy of the high frequency band. The results indicate that the mutual information is only a small fraction of the perceived entropy of the high-band. This suggests that speech-bandwidth extension should not rely only on mutual information between narrow- and high-band spectra. Rather, such methods need to make use of perceptual properties to ensure that the extended signal sounds pleasant.

During 2001 we continued our investigation of codebook driven speech parameter estimation. The work focused on the (short-term) linear predictor parameters, which describe the spectral envelope of the speech signal. Two pre-trained codebooks, one for speech spectral shapes, and one for noise spectral shapes, are utilized. For each pair of speech and noise spectra, the optimal mixing (for a given speech frame) is found and the corresponding likelihood is calculated. The pair of speech and noise spectra that maximizes the likelihood constitutes our estimate. This method provides robust estimates, since it is restricted to spectral shapes in the codebooks. The work was presented at ICASSP 2001. Further, by the end of 2001, the group received funding for two projects on noise suppression. To this end, two Ph.D. students and one post-doc were hired. The first project will focus on noise suppression for personal mobile communications, while the second will focus on professional (police, fire brigade, rescue services) mobile communications.

Auditory modeling

In speech and audio processing, it is important to understand the human perception of the signals. Improved understanding may lead to new quantitative distortion criteria and new coding algorithms. Our work focuses on two areas: the description and perceptual importance of phase in speech, which is an important topic for low-rate coding, and the development of a new coding paradigm where we code in the perceptual domain rather than the speech domain.

In our investigation of phase, we followed our studies on phase capacity, in which we evaluated information measures of the ability of the human auditory system to perceive phase. Our current study is directly relevant for the quantization of phase information performed in speech coders. Using two sophisticated auditory models, we investigated whether the squared error is a perceptually accurate measure for Fourier phase distortions in voiced speech-like signals. We considered the squared error between an original time-domain signal and the phase-distorted time-domain signal. The perceptual distortion was measured using the two auditory models. Two types of experiments were carried out: in the first type the direct

relationship between squared error and perceptual distortion was found, in the second type vector-quantization performance was investigated. It was found that low squared-errors correlate well to the perceptual error. For higher squared errors, a further increase in squared error does, on average, not lead to any further increase in perceptual error. This is consistent with the empirical rate perceptual-distortion curves we found. These curves show no decrease of perceptual distortion from low to medium rates for squared error based vector quantizers. Both auditory models gave consistent results, which we verified with a listening test. We showed that the observed behavior is a result of different sensitivity to time shifts in the human auditory system and the squared error. Empirical rate perceptual-distortion curves for a vector quantizer, using a perceptual criterion, showed that phase vectors are a particular undesirable parameter to quantize since they do not allow for efficient encoding. Motivated by these results we started investigating time-domain envelopes as a representation of phase spectra. We showed that time-domain envelopes within auditory channels are a perceptually correct representation while the time-domain envelope of the full-band speech signal is not.

The finding that the envelope representation requires the signal to be split into auditory channels supports our efforts in exploring new speech and audio-coding methods based on perception. For speech coders that fall within the class of waveform coders, the reconstructed signal approaches the original with increasing bit rate. In such coders, the distortion criterion generally operates on the speech signal or a signal obtained by adaptive linear filtering of the speech signal. To satisfy computational and delay constraints, the distortion criterion must be reduced to a very simple approximation of the auditory system. This drawback of conventional approaches motivates a new speech-coding paradigm in which the coding is performed in a domain where the single-letter squared-error criterion forms an accurate representation of perception. The new paradigm requires a model of the auditory periphery that is accurate, can be inverted with relatively low computational effort, and represents the signal with relatively few parameters. Our current results indicate that the new paradigm in general and our auditory model in particular form a promising basis for the coding of both speech and audio at low bit rates.

Voice and audio over the Internet

The properties of packet networks using the Internet Protocol (IP) differ significantly from those of the switched-circuit networks that were traditionally used for the transmission of voice and audio. The individual packets in which the coded information is contained can be lost or delayed, especially when traffic is close to or exceeds network capacity. If the end-to-end delay is of secondary importance, such as in radio broadcast, Automatic Repeat reQuest (ARQ), Forward Error Correction (FEC), and interleaving can be successfully used for packet-loss recovery. However, if a low-delay is of critical importance, then low-delay FEC and Multiple-Description Coding (MDC) schemes can be applied. We have focused on the MDC, where the decoding quality using any subset of the signal is acceptable, and better quality is obtained using more descriptions. From 2002, our research interests will focus more on robust source coding, adaptive error control coding and traffic control, jointly working with the Department of Signals, Sensors and Systems at KTH.

Filename: tsb.doc
Directory: I:\annual\Annual2001
Template: J:\tmh.kth.se\tmh\proj\common\template\tmh\Normal.dot
Title: Annual Report 2000
Subject:
Author: harald
Keywords:
Comments:
Creation Date: 2002-04-11 09:22
Change Number: 13
Last Saved On: 2002-05-22 16:24
Last Saved By: Personal
Total Editing Time: 86 Minutes
Last Printed On: 2002-06-12 10:40
As of Last Complete Printing
Number of Pages: 4
Number of Words: 2 056 (approx.)
Number of Characters: 11 721 (approx.)