SPEECH SIGNAL PROCESSING

Bastiaan Kleijn
Professor of
Speech Signal Processing

Speech signal processing

During the year 2000, the Speech Signal Processing Group at the Department of Speech Music and Hearing at KTH consisted of six Ph.D. students (four of whom were located at the department), a postdoc, a part-time (20%) researcher (forskarassistent), 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 telecommunication services and speech enhancement is used for communications in adverse environments. Despite these recent advances, lowering the bit rate and increasing quality remain important challenges for the future. For example, the required fixed bit-rate for toll quality is halved about every five years and there is no sign of a slowing of this rate of advance. In addition, new communication network technologies have introduced fresh challenges.

In the following, we provide a brief overview of the main research activities of the group during the year 2000.

Speech coding and synthesis

In speech coding and synthesis, the work of the group has continued its focus on the waveform interpolation (WI) algorithm. In addition, work was performed on the estimation of linear-prediction parameters.

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 which 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. We showed that by exploiting adaptive frames for the second expansion, problems in the voiced-unvoiced decomposition near onsets can, at least in principle, be eliminated. The scheme naturally leads to an adaptive bit allocation within the time-frequency plane.

Linear prediction (LP) is commonly used in speech processing, and is an integral part of most speech coding and synthesis systems. We have developed an improved LP method for which the associated all-pole model provides a better estimate of the vocal-tract transfer function than conventional LP methods. This should lead to, for example, more efficient coding and in improved pitch scaling for concatenative speech synthesis. In voiced speech, LP-based all-pole spectral envelopes often exhibit unnatural vocal tract transfer functions that underestimate the formant bandwidths. To obtain improved all-pole spectral envelopes we developed a regularization measure which discourages nonsmooth behavior of the transfer function. This regularization scheme can be incorporated into the LP framework without the need for iterative numerical optimization or spectral sampling. Our experimental results confirmed that regularized LP all-pole models can provide more accurate vocal tract transfer function modeling than conventional LP, particularly at the formants.

Audio coding

Traditionally, audio coders have used non-parametric 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 theNetherlands. 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. In each segment, the sinusoids are selected using a matching pursuit approach. To improve performance, a new matching pursuit algorithm which incorporates psychoacoustical modeling was developed. In this method, a psychoacoustic-adaptive norm on the signal space is defined, that can be used for selecting the dictionary elements in a rate-distortion optimal manner.

The sinusoidal model includes damped sinusoids. We increase the modeling efficiency of the damped sinusoids by modifying the locations of signal transients. As a result of the modifications, a transient will only occur at the beginning of a sinusoidal segment. In our first implementation of this concept, the transient component is subtracted from the signal and then added so that it is located at segment boundary. The modified signal is perceptually indistinguishable from the original signal.

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. We introduced a new aspect into the bandwidth extension procedure. For many cases of bandwidth extension, the high-band energy is over-estimated, leading to undesirable audible artifacts. To overcome these problems, we introduced an asymmetric cost-function in the estimation process of the high-band that penalizes over-estimates more than under-estimates of the energy in the high-band. We showed that the resulting attenuation of the estimated high-band energy depends on the broadness of the a-posteriori distribution of the energy given the extracted information about the narrow-band. Thus, the uncertainty about how to extend the signal at the high-band influences the level of extension. Results from listening tests show that the proposed algorithm produces less artifacts.

In our work on parameter estimation under environmental noise conditions, we focused on the (short-term) linear predictor parameters, which describe the spectral envelope of the speech signal. This work has practical relevance for, for example, speech coding and speech recognition in, for example, a car environment. We store separately the possible autoregressive spectral shapes of both the speech and the additive noise. The product codebook is then searched to maximize the likelihood function of the observed noisy speech signal frame. The Maximum Likelihood (ML) estimates of the variances of the driving term are computed for each pair of the speech and noise AR spectra. For further processing (e.g., Kalman filtering or speech coding using the enhanced linear-predictor parameters), the spectra and variances that yield the maximum of the likelihood function are selected. To evaluate the proposed method, the estimates of the spectral shapes and variances are compared with those computed from clean speech signal using a common spectral distortion measure. Globally maximizing the likelihood function over a restricted region of the parameter space, the presented approach provides robust estimates.

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 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, with a study which is directly relevant for the quantization of phase information performed in speech coders. Using a sophisticated auditory model, we investigated how accurately the squared error captures perceptual errors introduced by Fourier phase spectrum changes. We found that the squared error represents the perceptual error well for low squared errors but saturates for higher squared errors. This means a further increase in squared error does (on average) not lead to any further increase in perceptual error. This, in turn, suggests that encoding phase using squared-error trained codebooks only improves perceived quality when operating in the not saturated regions. To verify this, phase was encoded with codebooks of different sizes using the squared-error criterion during encoding. As expected, increasing the codebook size has very little influence on the average perceptual error for low rates, which was confirmed by listening tests. Our results suggest that a direct phase codebook is an unattractive representation of the relevant information contained in phase.

We are also exploring new speech and audio coding methods based on perception. For speech coders which 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 which 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 which 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, particular when traffic is close to or exceeds network capacity. New techniques must be developed to transmit speech and audio signals efficiently over these types of networks.

The Internet Protocol makes efficient use of network resources only when allowing routers to drop packets. Such packet loss causes degradation of perceived signal quality. To mitigate this degradation, advanced protocol mechanisms have been proposed as well as two classes of signal processing methods: a first class which adds redundancy at the transmitter side, e.g., loss-resilient codes or multiple-description source/channel coding, and a second class which performs packet-loss concealment through signal interpolation at the receiver side. We introduced a new class of signal processing methods which modify the transmitter without increasing the network payload data rate while minimizing the perceptual effect of packet losses at the receiver side. Our method operates in a situation where different IP packet streams travel over the same path between two gateways. Multiplexing can then be used to distribute a part of the information about one sound segment in several data packets while keeping the data packet rate and payload for that part of the information unchanged. This can be used to obtain packet loss characteristics which are less objectionable to the listener than those originating from a packet loss concealment method.