During the year 2002, the Speech Processing Group at KTH consisted of eight Ph.D. students, one post-doctoral researcher, and a professor. The group hosted two guest researchers. The group performs research encompassed within speech processing, signal processing, and source coding and teaches two undergraduate courses (Information Theory and Source Coding, and Speech Signal Processing), in addition to a varying number of graduate courses. The group also supervises numerous five-month projects performed by undergraduate students.

During 2002, the group was associated with the Department of Speech, Music and Hearing and co-located with the Department of Signals, Sensors, and Systems. (The group moved from the Dept. of Speech, Music and Hearing to the Dept. of Signals, Sensors, and Systems on January 1st, 2003.)

Research activities in the group focus mainly on developing algorithms for speech enhancement, audio and speech coding, source-coding techniques aimed at packet network environments, and speech synthesis. The state of the art in these areas is progressing rapidly. The outcome of the research that group members have actively participated in has become part of everyday life. Speech enhancement is becoming increasingly important for communications in adverse environments and also as a useful front end for speech recognition. Speech coding and synthesis form an enabling technology for mobile telephones. Audio coding is becoming commonplace in consumer electronic devices.

In the following, a brief overview of the main research activities of the group during the year 2002 is provided.

Speech Signal Processing

Speech Enhancement

Work on speech enhancement was carried out in three separate projects. Two projects focused on different aspects of single-channel speech enhancement and the third project focused on bandwidth extension.

The first project was aimed at the development of algorithms that exploit a-priori information about noise and speech in single-channel speech enhancement. The starting point of this work is formed by earlier work of the group that uses a codebook-constrained signal separation to instantaneously obtaining estimates of the speech and noise envelopes, and the signal power of the speech and noise. The method uses two codebooks, one codebook of speech envelopes and one codebook of noise envelopes (parameterized as line spectral frequencies). Given a speech frame, an envelope from the speech codebook, and an envelope from the noise codebook, the optimal speech and noise powers are estimated. This is referred to as local optimization. It is done for each pair of envelopes from the speech and noise codebook. In the subsequent global optimization, the optimal pair is chosen as the estimate of the speech and noise envelopes.

The method has the potential of estimating noise, and speech envelopes based on only one observed frame (typically 20 ms of length) of noisy speech. This distinguishes it from recently published state-of-the-art noise estimation techniques based on quantiles and minimum statistics, which do not perform well in non-stationary noise environments. Preliminary tests of the robustness of the method for signals corrupted by non-stationary noise are promising.

Work in the second project focused on the exploitation of perception in speech enhancement. Commonly used techniques in
speech enhancement are based on linear filters (Wiener and Kalman filters) that minimize the mean squared error. Motivated by the fact that minimization of the mean squared error is not always consistent with the human perception, possibilities for introducing perceptually-motivated modifications into these systems were investigated. It was found that performance could be improved significantly if the estimation and modeling accuracy of the enhancement system is considered. This project emphasizes methods that perform better in the presence of the non-stationary noise.

In the area of bandwidth extension, research was carried out on the estimation of mutual information between frequency bands in speech. The group 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 are 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, the dependency between bands through their mutual information and the perceived entropy of the high frequency band was quantified. 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.

**Speech Coding and Synthesis**

The group developed a new speech-coding paradigm based on a Karhunen-Loeve transform (KLT)-based adaptive classified vector quantizer (CVQ). In contrast to the commonly used CELP algorithm, the new method utilizes the space-filling advantage of vector quantization, since the shape of the Voronoi region is not affected by the KLT. The memory and shape advantages can be also used, since each codebook is designed based on a narrow class of KLT-domain statistics. The basic KLT-CVQ was further enhanced with companding. Experiments indicated that KLT-CVQ provides a higher SNR than the basic CELP coding structure, at a computational complexity similar to direct vector quantization (DVQ), which is much lower than CELP. With companding, even single-class KLT-CVQ outperformed CELP, both in terms of SNR and codebook search complexity.

Linear prediction is ubiquitous in speech coding (indeed, it forms part of the fore-mentioned KLT-CVQ based coder). In linear-prediction based speech coders, the predictor is generally encoded as side-information and interpolated between updates. It has been found that the line spectral frequency (LSF) representation of the predictor is particularly well suited for and interpolation. Despite the extensive use of the LSFs, insight in their behavior and meaning is less developed than that of other predictor representations, although some physical interpretations have been provided. Work together with Helsinki University of Technology in this area aims to provide additional understanding of the LSFs by showing that they can be determined from particular optimal constrained predictors. In particular, it was shown that, for even predictor order, the LSFs can be interpreted as the roots of two particular predictor polynomials. The corresponding predictors are optimal constrained predictors that correspond to prediction of the original signal from a particular high-pass and a particular low-pass filtered signal, respectively.

Enhancement of coded speech has been an important aspect in making the output of speech coders sound better. The group developed a new method for the enhancement of speech signals contaminated by speech-correlated noise, such as that in the output of a speech coder. For this work, the group won the Best Paper Award at the IEEE Speech Coding Workshop, 2002, Japan, for the paper “Enhancement of Coded Speech by Constrained Optimization”, by W.B. Kleijn. The method is based on constrained optimization of a criterion. The method was implemented on a block-by-block basis and use two constraints. A first constraint ensures that the signal power is preserved. A modification constraint ensures that the power of the difference of the enhanced and unenhanced signal is less than a fraction of the power of the unenhanced signal. The method was applied to increase the periodicity of the speech signal. Noisy-sounding coded voiced speech generally has a high SNR and can be enhanced while
satisfying a strict modification constraint. Sounds that are not nearly periodic are perceptually unaffected by the optimization because of the modification constraint. Practical results show that the method forms a powerful alternative to existing post-filter procedures.

**Audio Coding**

An important aspect of an audio coder is quantization of a signal representation. One of the commonly used signal models is the sinusoidal model, which operates on a segmental basis. In sinusoidal modeling, the signal representation includes amplitudes and phases of the sinusoidal components. The work was a continuation of earlier work of the group on unrestricted polar quantization where phase quantization accuracy depends on amplitude. New quantizers under the entropy constraint using high-rate assumptions were derived. Their distortion-rate performance for a circularly-symmetric (or bivariate) Gaussian input variable and a mean-squared error distortion measure was evaluated. It was shown that the theoretical performance of entropy-constrained unrestricted polar quantization is significantly higher than that of strictly polar quantization, where phase quantization is independent of amplitude, identical to rectangular quantization, and slightly inferior to two-dimensional vector quantization. Practical high-rate theory based unrestricted polar quantizers are shown to have performance very close to the theoretical prediction, even at low rates. The unrestricted polar quantization is generalized to the case of multiple sinusoids and a weighted error measure, which can account for the masking effect in the human auditory system. The amplitude and phase quantizers are defined to be dependent on the masking threshold. The new quantization method is used in an audio-coding application and is shown to significantly outperform a conventional sinusoidal quantization method where phase quantization accuracy is identical for all audible sinusoids.

The perceptual quality of audio coders can be improved by introducing distortion measures related to the human auditory system. Advanced perceptual distortion measures have a support length (the length of impact that one event in the input has on the distortion measure output) that exceeds typical coding block-lengths by far. To be able to code with these distortion measures it is necessary to either account for this effect in the coding method or to modify the distortion measure. Following the first approach, joint optimization of consecutive decisions (delayed decision coding) is a possible solution. The latter approach would involve evaluation of how these advanced perceptual distortion measures can be simplified and what the penalty is when a simplified measure is used.

**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. The work focuses on two areas: the description and perceptual importance of Fourier phase in speech, which is an important topic for low-rate coding, and the development of a new coding paradigm where coding is performed in the perceptual domain rather than the speech domain.

Based on two well-known auditory models it was investigated whether the squared error between an original signal and a phase-distorted signal is a perceptually relevant measure for distortions in the Fourier phase spectrum of periodic signals obtained from speech. Both the performance of phase vector quantizers and the direct relationship between the squared error and two perceptual distortion measures were studied. The results indicate that for small values the squared error correlates well to the perceptual measures. However, for large errors, an increase in squared error does, on average, not lead to an increase in the perceptual measures. Empirical rate-perceptual distortion curves and listening tests confirm that, for low to medium codebook sizes, the average perceived distortion does not decrease with increasing codebook size when the squared error is used as encoding criterion. The observed behavior is explained by the different sensitivity to local time shifts of the squared error and perceptual distortion measures.

**Source Coding for Packet Networks**

To facilitate audio-visual communications through a wireless Internet, packet-loss and bit-error-recovery techniques should be exploited. For real-time applications such as Voice over IP
where delay constraints are severe, both forward error correction (FEC) and multiple-description coding (MDC) are promising packet-loss recovery techniques.

The mean distortions associated with FEC and MDC in an environment with only packet losses from a rate-distortion viewpoint were compared. A one-dimensional Gaussian source with zero mean and unit variance was assumed. It was assumed that FEC was transmitted over a single channel and that two independent transmission channels were available for MDC. For MDC, we considered the balanced case only. The rate-distortion simulations show that FEC is robust against single-packet loss, while side-distortion optimized MDC is more robust against burst-packet loss. A channel-optimized MDC provides the same performance as the side-distortion optimized MDC for a high burst-packet-loss environment. However, channel-optimized MDC also provides a high SNR for the low burst-packet-loss case, thus clearly showing the advantage of informed coding. FEC always has a catastrophic failure (the packet can either be decoded or not), whereas in MDC we have a soft failure (for two descriptions, both descriptions of the packet can be decoded, or one description of packet can be decoded, or no information about the packet can be decoded). These differences in reconstruction behavior might have important perceptual repercussions that are not reflected in the mean squared-error distortion measure. Note that optimal FEC and MDC should give the same performance for identical redundancy levels for the error-free condition, if the MDC is optimized for the central channel.