Estimation of Sensory Information Transmission Using a Hidden Markov Model of Speech Stimuli

Arne Leijon
Dept. of Speech, Music and Hearing, KTH, Dr. Kristinas väg 31, SE–100 44 Stockholm, Sweden

Summary
A method is presented which gives good approximate estimates of the rate of information (in bits/s) successfully transmitted from a speech source to the modelled neural output of the peripheral sensory system. This information rate sets definite upper limits on the listener's speech-recognition performance. The performance limits depend on the entropy and vocabulary size of the speech material. The estimates of sensory information rate can be used to evaluate to what extent a listeners' performance is limited by peripheral loss of information or by suboptimal central processing. Calculations for a Swedish sentence test material, with an excitation-pattern auditory model, were consistent with human speech recognition results in speech-shaped masking noise. This suggests that the scarcity of sensory information may be the primary limiting factor in this test condition. Similar calculations for low-pass- and high-pass-filtered clean speech indicated a higher sensory information rate than required for the listeners' actual performance. These results suggest that the speech recognition performance under masking and filtering may be limited by different mechanisms. The analysis also showed that the information in adjacent frequency bands is not additive.

PACS no. 43.66.Ba, 43.71.An, 43.71.Cq

1. Introduction

The Articulation Index (AI) [1, 2], recently revised as the Speech Intelligibility Index (SII) [3], and the closely related Speech Transmission Index (STI) [4, 5, 6] have been quite successful in predicting speech intelligibility from objective acoustical measurements. The SII has also been applied to predict the effect of (quasi-linear) hearing instruments, using simple modifications to account for suprathreshold effects of sensorineural hearing loss [7, 8, 9]. For the purpose of choosing among hearing-aid frequency responses, SII-based estimates of speech intelligibility are actually more reliable than common speech recognition test results [10]. The success of these simple methods is quite astonishing in view of the vast complexity of human hearing and speech recognition.

These objective methods are based on the simplifying assumption that the contributions to speech intelligibility can be added across individual frequency bands. Although this model seems to work well in many situations, it has been shown that redundancy and synergy effects across frequency bands are significant under some filtering conditions [11, 12, 13, 14]. There are also some interactions between linguistic and acoustical factors that cannot be captured by a single objective index. Therefore, the SII uses different empirically determined importance functions for different types of speech material. Furthermore, it is not clear if these models can be modified to predict the effect of advanced hearing-aid signal processing, such as fast multichannel compression, noise reduction, or spectral enhancement. There are probably also individual variations in hearing ability, which cannot be predicted simply from hearing thresholds.

This paper presents an alternative approach to estimate the speech transmission capacity of a listener's sensory system in a given acoustical environment. The proposed method calculates the rate of speech-related information that is successfully transmitted from a speaker, in a noisy acoustic environment, through the listener's peripheral sensory system. This method does not model the speech recognition mechanism itself and, thus, cannot predict the actual performance of an individual listener, when tested with a specific speech material. It can, however, predict definite upper limits on the speech-recognition performance. Therefore, it can possibly predict the effects of hearing-aid signal processing on the *potential* for speech understanding in a given environment.

Previous versions of this approach applied some very crude approximations, because of the computational complexity [15, 16]. The method has now been improved to allow reasonably fast and accurate calculations, with few restrictions on the sensory model and on the non-linear hearing-aid signal processing. The specific purpose of this presentation is (A) to show the feasibility of the method, and (B) to analyse the information transmission under conditions of wideband masking and steep filtering, in order to shed some new light on early findings [11] that filter-

Received 10 September 2001, revised 19 November 2001, accepted 11 December 2001.

ing and masking may reduce speech intelligibility by different mechanisms. Another purpose (C) is to test the hypothesis [12, 13, 14] that speech variations in adjacent frequency bands are highly correlated, and therefore convey non-additive information, whereas the variations in widely separated bands are statistically independent, and therefore convey additive information contributions.

2. Theory and methods

2.1. Information and receiver performance

Information theory can not predict the actual performance of a real communication system, but it provides powerful tools to predict absolute upper limits on the possible performance, regardless of how the transmission system is implemented. The predictions of information theory are equally valid for human communication as well as for technical communication systems.

A block diagram of speech transmission from talker to listener is shown in Figure 1. It is assumed that a word sequence $\boldsymbol{W}$, when articulated, produces an acoustic signal $\boldsymbol{X}$, characterised e.g. by a sequence of short-time spectra $\boldsymbol{X}_t$, analysed with a fixed time resolution and sampling interval. The acoustic pattern sequence can also be characterised by a corresponding sequence $\boldsymbol{S}$ with class labels S_t, representing the phonetic category of each acoustic pattern. Figure 1 is not intended to claim that the word articulation process necessarily requires an intermediate phonetic coding step. The "phonetic" class label sequence can be seen simply as a convenient means to characterise the acoustic sequence. The acoustic signal is mixed with noise, possibly processed through a hearing aid, and then processed by the peripheral sensory system. The sensory pattern sequence $\boldsymbol{R}$ is then analysed by a central speech-recognition system to yield the received word message $\overline{\boldsymbol{W}}$. All pattern sequences are regarded as random sequences. The sequences $\boldsymbol{W}$, $\boldsymbol{S}$, and $\overline{\boldsymbol{W}}$ are discrete, with countable outcomes, whereas the vector sequences $\boldsymbol{X}$ and $\boldsymbol{R}$ are continuous-valued and must be described by probability density functions, denoted as e.g. $f_{\mathbf{R}}(\boldsymbol{r})$.

The amount of information transmitted successfully from one point to another in any communication chain, e.g. from $\boldsymbol{S}$ to $\boldsymbol{R}$, is quantified by the Mutual Information (MI), defined as

$$\begin{aligned} I(\boldsymbol{S};\boldsymbol{R}) &= H(\boldsymbol{S}) - H(\boldsymbol{S}|\boldsymbol{R}), \\ \text{where} \quad H(\boldsymbol{S}) &= E\big[-\log_2 P(\boldsymbol{S})\big] \\ &= -\sum_k P(\boldsymbol{S}=k)\log_2 P(\boldsymbol{S}=k), \\ H(\boldsymbol{S}|\boldsymbol{R}) &= E\big[-\log_2 P(\boldsymbol{S}|\boldsymbol{R})\big] \\ &= -\int_{\boldsymbol{r}} f_{\mathbf{R}}(\boldsymbol{r}) \sum_k P(\boldsymbol{S}=k|\boldsymbol{R}=\boldsymbol{r}) \\ &\qquad \cdot \log_2 P(\boldsymbol{S}=k|\boldsymbol{R}=\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}. \end{aligned}$$

Here and in the following, $E[\ ]$ denotes the expectation value. The integral over $\boldsymbol{r}$ and the sum over k are to be taken for all possible outcomes of $\boldsymbol{R}$ and $\boldsymbol{S}$. The entropy $H(\boldsymbol{S})$ measures the average *a priori* uncertainty about possible future messages $\boldsymbol{S}$, using knowledge only about the general statistical characteristics of $\boldsymbol{S}$. The conditional entropy $H(\boldsymbol{S}|\boldsymbol{R})$ measures the average *a posteriori* uncertainty about $\boldsymbol{S}$, given an observed sequence $\boldsymbol{R}$. Thus, $I(\boldsymbol{S};\boldsymbol{R})$ is a measure of the average reduction in uncertainty about $\boldsymbol{S}$, caused by observing $\boldsymbol{R}$. Reduction of uncertainty is obviously the same as a gain of information. Mutual information is a symmetric measure, and $I(\boldsymbol{S};\boldsymbol{R})$ is also a measure of the average reduction in the uncertainty about $\boldsymbol{R}$, caused by observing $\boldsymbol{S}$. Using the differential entropy

$$h(\boldsymbol{R}) = E\big[-\log_2 f_{\mathbf{R}}(\boldsymbol{R})\big],$$

the MI can also be expressed as

$$\begin{aligned} I(\boldsymbol{S};\boldsymbol{R}) &= h(\boldsymbol{R}) - h(\boldsymbol{R}|\boldsymbol{S}) = E\left[\log_2 \frac{f_{\mathbf{R}|S}(\boldsymbol{R}|\boldsymbol{S})}{f_{\mathbf{R}}(\boldsymbol{R})}\right] \\ &= \sum_k P(\boldsymbol{S}=k)\int_{\boldsymbol{r}} f_{\mathbf{R}|S}(\boldsymbol{r}|k)\frac{f_{\mathbf{R}|S}(\boldsymbol{r}|k)}{f_{\mathbf{R}}(\boldsymbol{r})}\,\mathrm{d}\boldsymbol{r}. \end{aligned}$$

Thus, the MI can also be seen as an average log-likelihood ratio for a signal classifier, designed to make optimal decisions about $\boldsymbol{S}$, given observations $\boldsymbol{R}$.

The MI is an interesting quantity for several reasons. The information transmitted from end to end of a communication chain cannot exceed the amount of transmitted information through each link in the chain. The so-called data processing inequality [17] implies that

$$I(\boldsymbol{W};\overline{\boldsymbol{W}}) \le I(\boldsymbol{W};\boldsymbol{R}) \le I(\boldsymbol{S};\boldsymbol{R}).$$

Furthermore, the mutual information $I(\boldsymbol{W};\overline{\boldsymbol{W}})$ sets a definite upper bound on the performance of the system, quantified for example as the probability of correct recognition, $P_c = P(\overline{\boldsymbol{W}} = \boldsymbol{W})$. The relation between P_c and $I(\boldsymbol{W};\overline{\boldsymbol{W}})$ depends on the number of available response

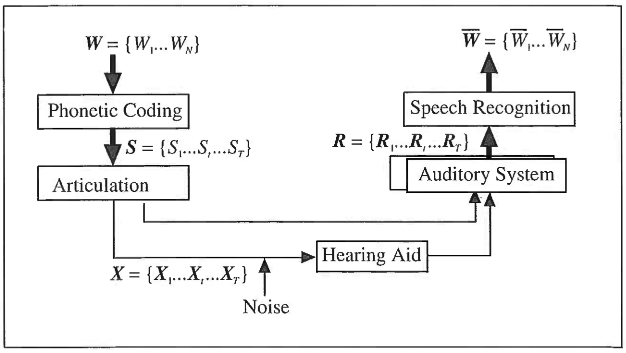

Figure 1. Block diagram of the speech communication chain, defining the notation used in this work. $\boldsymbol{W}$ is a transmitted sequence of N words, $\boldsymbol{X}$ is a sequence of short-time spectra representing the clean speech signal with a duration of T signal frames, $\boldsymbol{S}$ is a corresponding sequence of phonetic categories, and $\boldsymbol{R}$ is the resulting sequence of (possibly multimodal) sensory patterns.

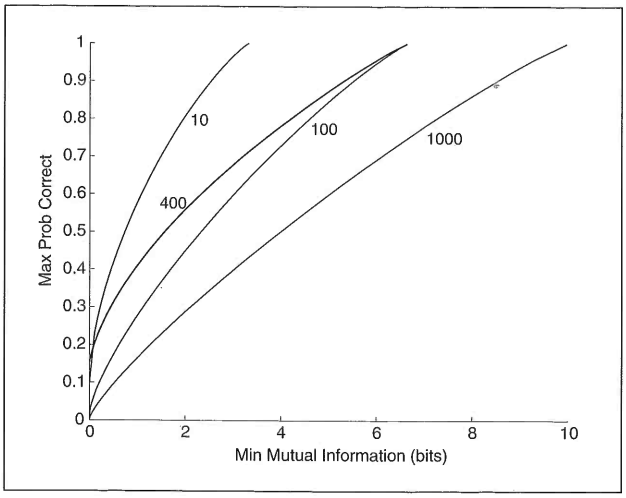

Figure 2. Rate-distortion functions, defining upper bounds on the probability of correct word recognition and corresponding lower bounds on the amount of successfully transmitted information (bits/word), for word test materials with different linguistic complexity. Curve parameters indicate the number of possible response alternatives. Thin lines represent tests with equal word probabilities, and the thicker line is valid for a word material with a more realistic "Zipf" distribution among 400 possible response alternatives (see text).

categories, e.g. word alternatives in a word-recognition test, and on the entropy of the distribution of the possible word alternatives. This so-called rate-distortion function can be calculated with the Blahut-Arimoto algorithm [17], and a few examples are illustrated in Figure 2. For any given amount of received information per test word, recognition improves with decreasing difficulty of the test, i.e. decreasing word entropy and/or decreasing number of possible confusions. Perfect word recognition, $P_c = 1$, is theoretically possible whenever the available information exceeds the entropy of the word distribution.

With a given number of word alternatives, the word entropy is maximal if all words occur with equal probability. Of course, this is not the case in real languages, where words tend to occur, according to "Zipf's Law", with probabilities approximately proportional to $1/n$, where n is the rank order of the word probability [18, 19]. Figure 2 shows an example of two word sets with equal entropy; one set of 100 words with uniform distribution, and one set of 400 words with a "Zipf" distribution. These word sets allow maximum recognition performance at the same amount of information. However, at lower performance levels the relation between information and word-recognition probability differs between the two sets.

The fixed relation between information and optimal recognition performance is reminiscent of the well-known relation between performance and detection index d' in an orthogonal M-alternative forced-choice task, where each response alternative is equally probable. This relation can model the effect of the number of response alternatives in formal word-recognition tests [13, 14, 20]. However, the rate-distortion relation exemplified in Figure 2 is more general, as it can be applied to any probability distribution of test units.

It is now proposed that the mutual information $I(\boldsymbol{S}; \boldsymbol{R})$ is the most interesting quantity to estimate as a measure of sensory information transmission. This quantity focuses on the peripheral sensory process and does not depend on the used word material nor on the listener's central speech recognition ability. Nevertheless, it sets definite upper limits on the listener's performance. The total mutual information, measured in bits, increases with the duration of the speech sequence. For an ongoing communication process, such as human speech, it is more convenient to use the *rate of mutual information*, measured in bits/frame, bits/s, or bits/word.

It is very difficult to calculate the rate of mutual information exactly. However, it is possible to derive approximate lower and upper bounds on the rate of mutual information, if a hidden Markov model (HMM) is used to represent the statistical properties of the sensory pattern sequence. Hidden Markov models have a simple mathematical structure, but they can, nevertheless, describe very complicated non- stationary signal patterns, such as speech. They are routinely used in automatic speech recognition work.

2.2. Derivation

This section describes a method to derive approximate lower and upper bounds of the mutual information rate between a discrete phonetic sequence in a spoken message and the corresponding stream of sensory patterns. The bounds are estimated by a Monte Carlo technique, using a hidden Markov model to describe the statistical properties of the sensory data. In principle, it is a straightforward application of information theory to derive the desired lower and upper bounds, when the hidden Markov model is known. The main challenge is to estimate the parameters of the model and to obtain numerical results with a reasonable computational effort.

The proposed method does not require a phonetically labelled input speech signal. Instead, the "phonetic" categories in the signal are derived automatically by training a HMM using the clean speech signal. Therefore, the method requires a recorded input signal with speech and noise stored in separate tracks. A special method is then used to adapt the model to include the effects of external noise and internal sensory noise.

2.2.1. Rate of mutual information

The MI rate in bits/frame is defined as

$$r = \lim_{t\to\infty} \frac{1}{t} I(S_1 \ldots S_t; \boldsymbol{R}_1 \ldots \boldsymbol{R}_t).$$

For a stationary and causal system this is equivalent to

$$r = \lim_{t\to\infty} \Big(I(S_1 \ldots S_t; \boldsymbol{R}_1 \ldots \boldsymbol{R}_t) \\ - I(S_1 \ldots S_{t-1}; \boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1}) \Big)$$

$$= \lim_{t\to\infty} \Big(h(\boldsymbol{R}_t|\boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1}) \\ - h(\boldsymbol{R}_t|S_1 \ldots S_t, \boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1}) \Big),$$

where

$$h(\boldsymbol{R}_t|\boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1}) = E\big[-\log_2 f(\boldsymbol{R}_t|\boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1})\big],$$

etc. For a stationary hidden Markov system, where each $\boldsymbol{R}_t$ depends statistically only on the corresponding S_t, the MI rate is simply

$$r = \lim_{t\to\infty} h(\boldsymbol{R}_t|\boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1}) - h(\boldsymbol{R}_t|S_t),$$

where $h(\boldsymbol{R}_t|S_t)$ is independent of t, as the process is assumed to be stationary. Tight upper and lower bounds on this MI rate can be calculated using a finite length D of the conditioning sequence:

$$r_{\text{low}}(D) \le r \le r_{\text{high}}(D),$$
$$r_{\text{low}}(D) = h(\boldsymbol{R}_{D+1}|S_1\boldsymbol{R}_2 \ldots \boldsymbol{R}_D) - h(\boldsymbol{R}_t|S_t),$$
$$r_{\text{high}}(D) = h(\boldsymbol{R}_{D+1}|\boldsymbol{R}_1\boldsymbol{R}_2 \ldots \boldsymbol{R}_D) - h(\boldsymbol{R}_t|S_t),$$

because of a general theorem for hidden Markov systems [17],

$$h(\boldsymbol{R}_{D+1}|S_1\boldsymbol{R}_2 \ldots \boldsymbol{R}_D) \le \lim_{t\to\infty} h(\boldsymbol{R}_t|\boldsymbol{R}_1 \ldots \boldsymbol{R}_{t-1}) \\ \le h(\boldsymbol{R}_{D+1}|\boldsymbol{R}_1\boldsymbol{R}_2 \ldots \boldsymbol{R}_D).$$

Although the sensory patterns are modelled as continuous random vector sequences, good estimates of the MI rate can be obtained using a quantised version $\boldsymbol{R}^q$, because $I(\boldsymbol{S};\boldsymbol{R}) \approx I(\boldsymbol{S};\boldsymbol{R}^q)$. The approximation error decreases towards zero with finer quantisation [17].

2.2.2. Hidden Markov model of noisy sensory data

The sensory HMM is developed in 9 steps. First, the clean speech signal is analysed spectrally and the sequence of short-time spectra is described by a discrete Hidden Markov model, adapted to this pattern sequence. The source states of this HMM are assumed to represent the most significant "phonetic" categories in the signal. Then, the speech and noise are mixed and processed through the sensory model. The resulting sensory pattern sequence is analysed in the same time-scale as the original clean speech. The sensory pattern sequence is described by another discrete hidden Markov model, using the same state sequence as in the first HMM. This model is then adapted to represent the noisy sensory data instead of the clean speech. The resulting HMM is then used to estimate the desired bounds on the rate of mutual information between source states and corresponding sensory patterns.

1. The clean speech signal is analysed into a sequence of vectors $\{\boldsymbol{X}_1 \ldots \boldsymbol{X}_t \ldots \boldsymbol{X}_T\}$ containing log-magnitude short-time spectra with uniform auditory ERB frequency resolution. The present implementation used a uniform sampling interval (= frame duration) of 23 ms and a frequency resolution of 0.5 ERB.

2. A Vector Quantizer codebook $C_X = \{\boldsymbol{e}_1 \ldots \boldsymbol{e}_L\}$ is created to represent the spectrum sequence data with minimum distortion, using the standard generalised Lloyd algorithm [21]. Using this codebook, the spectrum sequence can be represented by a corresponding sequence of integer codebook indices $\{l_1 \ldots l_t \ldots l_T\}$, where $\boldsymbol{X}_t \approx \boldsymbol{e}_{l_t}$. More precisely, $\boldsymbol{X}_t \in V(C_X, l_t)$, where $V(C_X, l)$ is the Voronoi region around the code vector $\boldsymbol{e}_l$, i.e. the region of points closer to $\boldsymbol{e}_l$ than to any other codebook entry. The present implementation used a codebook with $L = 200$ code vectors.

3. An ergodic discrete HMM is trained to represent the vector-quantized version of the spectrum sequence, using the standard Baum-Welch algorithm [22]. (The present implementation used a model with 20 states.) This HMM represents the clean speech signal and is defined by the triplet $\lambda_a = \{A, B, \boldsymbol{p}_0\}$, where
A = state transition probability matrix with elements $a_{ij} = P(S_t = j|S_{t-1} = i)$,
B = observation probability matrix with elements $b_{jk} = P(\boldsymbol{X}_t \in V(C_X, k)|S_t = j)$,
$\boldsymbol{p}_0$ = initial state probability vector with elements $p_0(i) = P(S_1 = i)$.

4. The speech and noise channels are mixed at the desired signal/noise ratio and, possibly, processed to simulate e.g. a hearing aid with any type of linear or nonlinear characteristics. The resulting signal is then processed though a sensory model, using exactly the same time sampling interval as in step 1. The model may also include multimodal sensory input. (The auditory model used in the present study is described in section 2.3.) The model output is a sequence of vectors $\{\boldsymbol{R}_1 \ldots \boldsymbol{R}_t \ldots \boldsymbol{R}_T\}$ representing the sensory pattern sequence.

5. A new codebook $C_R = \{\boldsymbol{c}_1 \ldots \boldsymbol{c}_K\}$ is created, in the same way as in step 2, to represent the sequence of sensory data as $\{k_1 \ldots k_t \ldots k_T\}$, where $\boldsymbol{R}_t \approx \boldsymbol{c}_{k_t}$. The present implementation used a codebook of size $K = 200$.

6. A new ergodic discrete HMM is now created to represent the actually observed sensory pattern sequence as $\{A, G, \boldsymbol{p}_0\}$. Here, only the observation probability matrix differs from the HMM in step 3, as G must here represent the quantised sensory sequence $\{k_1 \ldots k_t \ldots k_T\}$ using codebook C_R, instead of the quantised clean-speech data $\{l_1 \ldots l_t \ldots l_T\}$ using codebook C_X. The standard HMM training procedure uses the forward-backward algorithm to estimate

$$\gamma_t(j) = P(S_t = j|l_1 \ldots l_T, A, B, \boldsymbol{p}_0).$$

Using this result, the observation probabilities in the G matrix are now estimated as

$$g_{jm} = \frac{\sum_{t\in\tau_m} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)},$$

where τ_m is the set of time indices where $k_t = m$.

7. The sensory data are necessarily noisy, because of neural Poisson-type variability or any other source of randomness, as defined by the sensory model. To simplify cal-

culations, the sensory data should be transformed so that the sensory noise can reasonably be assumed to be Gaussian with zero mean. The sensory model must allow the estimation of a covariance matrix for the sensory noise, given any input signal. In particular, the model must supply an estimated covariance matrix $\boldsymbol{K}_m$, assumed to be approximately valid for any $\boldsymbol{R} \approx \boldsymbol{c}_m$, for each code index m in the codebook. If the sensory noise signals are assumed to be uncorrelated among sensory "channels", these covariance matrices are diagonal. (In a very simple sensory model, the matrices may be identical, regardless of sensory input.) Now, using the Gaussian approximation, if a particular input signal causes random sensory responses with mean $\boldsymbol{c}_m$, the conditional probability density of observing any sensory response $\boldsymbol{r}$ is, with N-dimensional sensory vectors:

$$f_{\mathbf{R}}(\boldsymbol{r}|\boldsymbol{c}_m) = \frac{1}{(2\pi)^{N/2}\sqrt{|\boldsymbol{K}|_m}} \mathrm{e}^{-\frac{1}{2}(\mathbf{r}-\mathbf{c}_m)'\mathbf{K}_m^{-1}(\mathbf{r}-\mathbf{c}_m)}.$$

Of course, the input signals causing sensory patterns quantised as $\boldsymbol{c}_m$ do not all yield exactly a response $= \boldsymbol{c}_m$. The quantisation noise is approximated by a diagonal covariance matrix $\boldsymbol{Q}_m$, estimated during the VQ design. (In the present application the quantisation variance was actually much smaller than the sensory variance.)

8. The information loss caused by sensory noise can now be modelled. We transform the observation probability matrix to account for the probability of confusions between sensory observations. Using the HMM obtained in step 6 in combination with the sensory-noise covariance from step 7, we can estimate the continuous conditional probability density for observing any sensory pattern $\boldsymbol{r}$, given any source state j, as a Gaussian mixture density

$$f_{\mathbf{R}_t|S_t}(\boldsymbol{r}|j) \approx c \sum_{m=1}^{K} \frac{g_{jm}}{\sqrt{|\boldsymbol{K}_m + \boldsymbol{Q}_m|}} \cdot \mathrm{e}^{-\frac{1}{2}(\mathbf{r}-\mathbf{c}_m)'(\mathbf{K}_m+\boldsymbol{Q}_m)^{-1}(\mathbf{r}-\mathbf{c}_m)}.$$

The scaling constant c will be cancelled out later. These probability densities are now sampled at all codebook points $\boldsymbol{r} = \boldsymbol{c}_k$, and properly normalised to yield the conditional discrete probabilities

$$q_{jk} = \frac{f_{\mathbf{R}_t|S_t}(\boldsymbol{c}_k|j)}{\sum_{n=1}^{K} f_{\mathbf{R}_t|S_t}(\boldsymbol{c}_n|j)} \approx P(\boldsymbol{R}_t \in V(C_R, k)|S_t = j).$$

Collecting these conditional probabilities in a matrix Q, we now have a third ergodic discrete HMM defined by $\{A, Q, \boldsymbol{p}_0\}$. This model represents the random sensory pattern sequences that may be caused by the given input signal, including the random confusions caused by sensory noise.

9. Finally, the HMM should not describe the particular utterance at the beginning of the training material, but rather the average statistical characteristics of all the possible signals exemplified by the recording. In an ergodic HMM the state probability distribution tends towards a unique time-independent stationary distribution $\boldsymbol{p}_s$ as $t \to \infty$, regardless of the particular initial distribution $\boldsymbol{p}_0$. The stationary distribution is estimated simply as $\boldsymbol{p}_s = A'^N \boldsymbol{u}$, for a large N and $\boldsymbol{u}$ = a uniform distribution. The final model $\lambda_s = \{A, Q, \boldsymbol{p}_s\}$ is then used for calculating the average mutual information rate between model states and sensory patterns, as described in the next section.

2.2.3. Estimating mutual information bounds

Once the discrete HMM $\lambda_s = \{A, Q, \boldsymbol{p}_s\}$ for the sensory pattern stream is available, it is fairly straightforward to obtain approximate lower and upper bounds on the mutual information rate between model states S_t and corresponding quantised sensory pattern $\boldsymbol{R}_t$. For simplicity we now denote the discretized sensory pattern by Z_t, defined as $Z_t = k \Leftrightarrow \boldsymbol{R}_t \in V(C_R, k)$. The information-rate bounds are then

$$r_{\mathrm{low}}(D) = H(Z_{D+1}|S_1 Z_2 \ldots Z_D) - H(Z_t|S_t),$$
$$r_{\mathrm{high}}(D) = H(Z_{D+1}|Z_1 Z_2 \ldots Z_D) - H(Z_t|S_t).$$

Three entropy calculations are obviously required. The entropy $H(Z)$ for any discrete random variable Z, characterised by a probability distribution vector $\boldsymbol{p}$ with elements $p(i)$, is calculated as

$$H(Z) = e(\boldsymbol{p}) = -\sum_k p(k) \log_2 p(k).$$

As the conditional distribution $P(Z_t|S_t)$ is independent of t and given directly by the observation probability matrix Q we easily obtain

$$H(Z_t|S_t) = \sum_j e(\boldsymbol{q}_j) p_S(j),$$

where $\boldsymbol{q}_j$ are rows of the matrix Q, with elements q_{jk} and $p_S(j)$ are elements of the known state probability vector $\boldsymbol{p}_s$.

A Monte Carlo technique is used for the remaining two entropies. The model λ_s is used repeatedly to generate random state sequences $\{i_1 \ldots i_D\}$ and corresponding observation sequences $\{k_1 \ldots k_D\}$.

For each generated sequence, the standard forward algorithm [22] is used twice. First the forward algorithm is initialised with the stationary initial state probabilities, i.e. $P(S_1 = j) = p_S(j)$. The algorithm is iterated D steps forward to yield the conditional state probability vector $\boldsymbol{\alpha}_D^{\mathrm{high}}$, with elements

$$\alpha_D^{\mathrm{high}}(j) = P(S_D = j|Z_1 = k_1, \ldots Z_D = k_D).$$

The algorithm is then reinitialised with exact knowledge about the first state, i.e. $P(S_1 = i_1) = 1$, yielding a slightly different conditional state probability vector $\boldsymbol{\alpha}_D^{\mathrm{low}}$, with elements

$$\alpha_D^{\mathrm{low}}(j) = P(S_D = j|S_1 = i_1, Z_2 = k_2, \ldots Z_D = k_D).$$

The desired conditional probability distributions for Z_{D+1} are then easily obtained as $\boldsymbol{p}_{D+1}^{\text{high}} = Q'A'\boldsymbol{\alpha}_D^{\text{high}}$, with elements

$$p_{D+1}^{\text{high}}(k) = P(Z_{D+1} = k | Z_1 = k_1, \ldots Z_D = k_D),$$

and similarly for $\boldsymbol{p}_{D+1}^{\text{low}}$.

The resulting conditional entropies

$$e\left(\boldsymbol{p}_{D+1}^{\text{low}}\right) \quad \text{and} \quad e\left(\boldsymbol{p}_{D+1}^{\text{high}}\right)$$

are calculated and accumulated. The whole procedure is repeated over and over again, and the total mean entropy values are calculated over all Monte-Carlo replications of this procedure.

The obtained bounds $r_{\text{low}}(D)$ and $r_{\text{high}}(D)$ get tighter with increasing length D of the generated random sequences, and the estimated mean values get more stable with increasing number of replications. The present implementation used $D = 15$ for each generated sequence, and 10000 Monte-Carlo replications.

2.3. Auditory model

The previous section defined the computations required to obtain upper and lower bounds on the desired rate of mutual information. The method requires a model of sensory signal processing, yielding (A) a sensory pattern vector for each input signal frame, and (B) an estimate of the sensory-noise covariance matrix for any pattern vector.

For the present implementation the auditory processing was represented entirely in the frequency domain, yielding one auditory excitation-level pattern for each signal frame. The acoustic signal was first hanning-windowed and analysed by FFT giving a sequence of short-time power spectra with 23 ms time resolution. The signal was assumed to be presented in a diffuse sound field, and the spectra were therefore first weighted by an approximate transfer function from diffuse field to the eardrum, obtained as an average for all horizontal wave directions [23].

Cochlear filtering was modelled by a roex(p, w, t) filter using a non-linear combination of a "tail" filter and a "peak" filter [24, 25]. The non-linearity is output-controlled, i.e. the peak filter gain is controlled by the total power in the passband of the peak filter. The system is linear at high levels and compressive at lower levels. The tail and peak filters at place coordinate z give two separate excitation components by weighting the short-time power spectrum $S(f)$ over all frequencies between zero and half the sampling rate $(F_s/2)$:

$$E_{\text{tail}}(z) = \int_0^{F_s/2} W_{\text{tail}}(f, z) S(f)\,\mathrm{d}f,$$

$$E_{\text{peak}}(z) = \int_0^{F_s/2} W_{\text{peak}}(f, z) S(f)\,\mathrm{d}f,$$

where

$$W_{\text{peak}}(f, z) = \left(1 + p(z)\frac{|f - f_c(z)|}{f_c(z)}\right) \mathrm{e}^{-p(z)|f - f_c(z)|/f_c(z)},$$

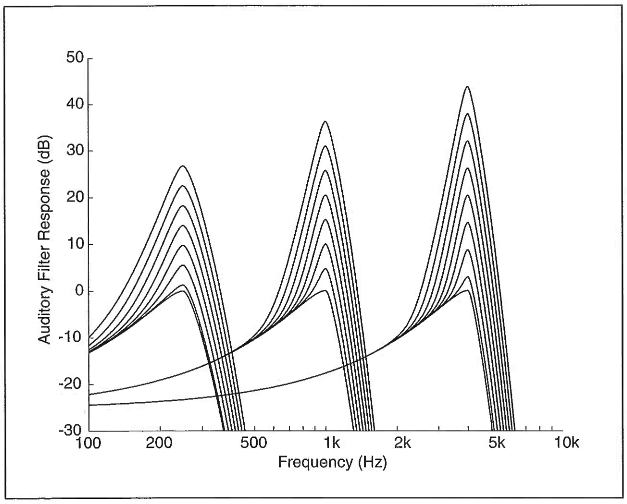

Figure 3. Examples of compressive magnitude frequency responses for modelled auditory filters at three centre frequencies. The peak gain is determined by the output from the peak filter. Responses are shown for filter output levels of 10, 20,. . . ,90 dB SPL at the eardrum.

$$W_{\text{tail}}(f, z) = \begin{cases} \left(1 + t\dfrac{f_c(z) - f}{f_c(z)}\right) \mathrm{e}^{-t(f_c(z) - f)/f_c(z)}, \\ \qquad\qquad\qquad\qquad f \le f_c(z), \\ \left(1 + p(z)\dfrac{f - f_c(z)}{f_c(z)}\right) \mathrm{e}^{-p(z)(f - f_c(z))/f_c(z)}, \\ \qquad\qquad\qquad\qquad f > f_c(z). \end{cases}$$

Here the parameter $p(z)$, defining the symmetric peak filter shape, was determined to yield a normal auditory ERB [26], and the ERB-rate frequency scale $f_c(z)$ was also defined by this relation. The t parameter defining the low-frequency slope of the tail filter was set to $t = 8$ [24]. The total excitation was then determined as a combination of the peak and tail components, plus a fixed excitation to represent the absolute threshold:

$$\begin{aligned} E(z) = {} & G\big(z, E_{\text{peak}}(z)\big) E_{\text{peak}}(z) \\ & + E_{\text{tail}}(z) + E_{\text{threshold}}(z), \end{aligned}$$

where

$$G\big(z, E_{\text{peak}}(z)\big) = \max\left(0, \frac{k(z)^{1-c(z)}}{E_{\text{peak}}(z)^{1-c(z)}} - 1\right).$$

The peak-excitation gain is zero when $E_{\text{peak}}(z)$ exceeds $k(z)$. Below this kneepoint, the total excitation becomes asymptotically proportional to $E_{\text{peak}}(z)^{c(z)}$, where $c(z) < 1$ defines the degree of compression. The parameters $k(z)$ and $c(z)$ were determined according to Figures 3 and 4 in [24]. The total non-linear filtering effect is illustrated in Figure 3. Finally, the model output was transformed to a logarithmic scale, $R(z) = 10\log_{10}(E(z))$, in order to allow approximately level-independent variance of the sensory noise. In the present implementation these excitation-level patterns were determined with a uniform place resolution of $\Delta z = 0.5$.

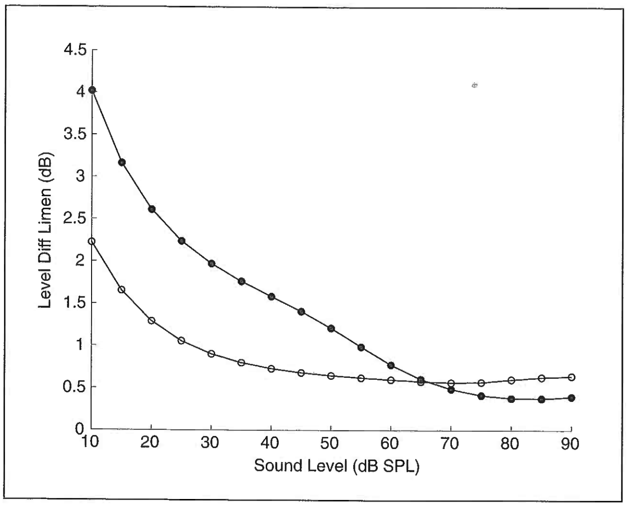

Figure 4. Intensity discrimination for white noise (open symbols) and for a pure tone at 1000 Hz (filled symbols), as predicted by the used auditory model for stimuli with 200 ms duration. The indicated level difference thresholds represent a detection index of $d' = 1$.

The internal sensory noise was assumed to be uncorrelated between z-scale channels. The model variance was determined to yield intensity discrimination for white noise and for tones as shown in Figure 4, which is in good agreement with empirical data [27, 28].

2.4. Speech material

Calculations were performed using a standardised Swedish word-recognition test material with known normal recognition scores in noise [29]. Each list consists of ten five-word "sentences" with identical syntactic structure and exactly 10 possible word alternatives at each position in the sentence. All lists are equivalent, as exactly the same 50 recorded words are used in each list. The noise masker has the same long-time spectrum as the speech material. All the calculations were done using one complete list where the pauses between sentences had been removed.

This speech material is obviously rather unnatural, but it has a definite advantage for this preliminary study: Its statistical characteristics are so simple that it is easy to apply rate-distortion theory to predict limits on the absolute recognition score from the calculated rate of mutual information (Figure 2). For a more natural speech material the information rates can be used only as relative measures, when comparing transmission systems.

3. Results

The sensory information rate for the chosen speech material was first calculated as a function of signal/noise ratio. The information rate is shown in Figure 5a. The corresponding theoretical upper bound on word recognition agrees fairly well with the empirical performance of listeners with normal hearing, as shown in Figure 5b.

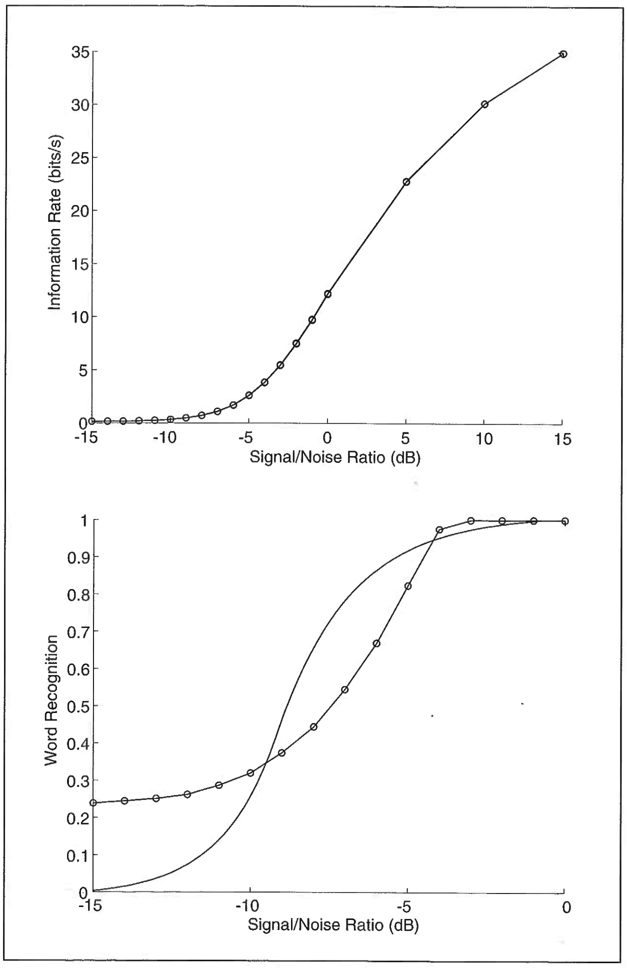

Figure 5. Calculated sensory information rate and corresponding limits on word recognition scores as a function of signal/noise ratio, estimated for a standardised Swedish word-test material and listeners with normal hearing. The speech level was fixed at 70 dB SPL and the noise had the same long-term spectrum shape as the speech. Panel A shows upper and lower bounds on the information rate in two indistinguishable curves. Panel B shows corresponding upper bounds on the probability of correct responses in this test material with exactly 10 equally probable response alternatives for each test word. The curve without symbols shows average empirical test results for normal listeners.

The effect of overall speech presentation level was small. The information rate was 91.9 bit/s with speech / noise presented at 52/57 dB SPL and 2.5 bit/s at 65/70 dB SPL. This difference corresponds to a difference of about 0.1 in predicted speech recognition scores. This difference is in opposite direction to the results of Hagerman [29], who found slightly better results at 52 dB SPL than at 65 dB SPL.

The information rate was also calculated for low-pass- and high-pass-filtered speech at a signal/noise ratio of 63/48 dB SPL. A series of cut-off frequencies were chosen to give stepwise increasing values of the Speech Intelligibility Index (SII) [3]. For comparison, the information rate was also calculated for wideband presentation with signal/noise ratios increasing from −15 to +15 dB in 3-dB steps, i.e. in steps of 0.1 SII units. The result, shown in

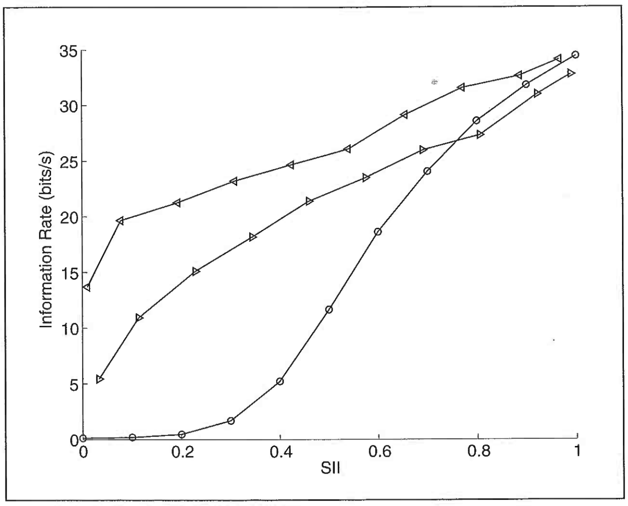

Figure 6. Sensory information rate, estimated for a standardised Swedish word-test material presented for listeners with normal hearing, varying the Speech Intelligibility Index (SII) by three methods: (1) Masked broadband presentation with varying signal/noise ratios (circles), (2) low-pass filtering with fixed signal/noise ratio at +15 dB (left-pointing triangles), and (3) high-pass filtering with signal/noise ratio at +15 dB (right-pointing triangles). Speech was always presented at 63 dB SPL.

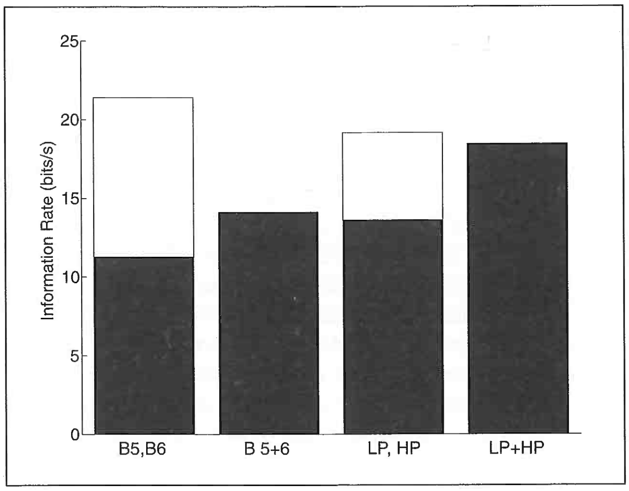

Figure 7. Sensory information rate, estimated for a standardised Swedish word-test material, filtered in standardised SII bands [3] 510–630 Hz (B5, first bar, black), 630–770 Hz (B6, first bar, white), 510–770 Hz (band 5+6 combined, second bar), 0–200 Hz (LP, third bar, black), 6.4–10 kHz (HP, white), and in the LP and HP bands combined (fourth bar). The speech/noise levels were 63/48 dB SPL.

Figure 6, indicates much higher information rates for the filtered conditions than for the masked presentations. The normal signal/noise threshold for 50% recognition in the masked broadband condition corresponds to SII≈0.2, but a low-pass-filter with cut-off at 200 Hz, corresponding to SII=0.01, gives an information rate of 13 bit/s, which is theoretically more than sufficient for 100% recognition.

The sensory information rate was also calculated for bandpass filtered speech, and for low-pass filtered and high-pass filtered speech, as well as for the sum of low-pass and high-pass-filtered speech. The bands were selected from Table I in the SII standard [3]. The results in Figure 7 show that bands 5 and 6 each conveyed about 10–12 bits/s, but bands 5+6 together conveyed only about 14 bits/s. The information rate was 13 bit/s in the low-pass band below 200 Hz, about 6 bit/s in in the high-pass band above 6400 Hz, and 18 bits/s in these bands combined. Thus, the information in adjacent bands is not additive, probably because the speech variations are correlated, whereas the information in widely separated bands seems to be almost completely additive.

4. Discussion

The present approach is similar to the Speech Intelligibility Index (SII) and the Speech Transmission Index (STI), in the sense that it attempts to derive a single number representing the speech-transmission capacity of the sensory system, without modelling the actual speech-recognition processes in the listener's brain. Another approach [30] is to use the sensory pattern sequence as input to an optimal decision mechanism that predicts the actual word-recognition performance. That approach is obviously appealing when the speech material has a limited number of response alternatives. However, as the present method does not require a speech-recognition model, it can be more easily applied to natural speech.

The consistency (Figure 5b) between model predictions and real listeners' performance under broadband masking conditions is quite astonishing, considering that the model calculation used no free fitting parameters. This result suggests that the scarcity of sensory information may be the primary limiting factor in this test condition. The rate of sensory information is an interesting measure of sensory performance, because it is independent of the linguistic entropy of the speech material and independent of the listener's linguistic skill.

Theoretically, no listener should be able to do better than the model-predicted upper bound. However, the simple auditory model used here discards some signal features, such as the voiced/unvoiced distinction and the voice pitch, which may be utilised by real listeners.

At very low signal/noise ratios the information rate is extremely low, but fig 5b shows that the predicted upper bound on performance is still clearly higher than chance level (0.1). This happens because an ideal classification system requires only very little information to perform better than chance, according to rate-distortion theory (Figure 2). Real listeners may give up when so little sensory information is available that the recognition task seems almost impossible.

The results for low-pass- and high-pass-filtered speech (Figure 6) imply that the filtered conditions yield more than sufficiently many distinguishable sensory patterns, although normal users are not able to utilise these pattern differences for speech recognition. There is a clear discrepancy between the results for broadband masking and filtering. This suggests that the speech recognition performance under masking and filtering may be limited by quite different mechanisms. The main limiting factor in broadband masking is probably the external noise. In the present model the internal sensory noise is the only factor that limits the result in the filtering condition. It might have been possible to obtain better consistency by increasing the variance of the sensory noise. However, then the auditory model would no longer be consistent with known level-discrimination data (Figure 4). For example, the 200 Hz low-pass filter result indicates that there were many distinguishable signal levels in the filtered stimulus and that these level variations were statistically correlated with the assigned phonetic classes. These level variations could convey speech information because the signal was presented at a fixed overall level that was known by the analysis model. Future work will investigate ways to prevent the model from utilising level variations that are disregarded by human listeners.

The results (Figure 7) also support the hypothesis that information in adjacent critical bands is not independent. Studies on phonetic matching of two-formant and one-formant synthetic vowels have indicated spectral integration over a range of 3–3.5 Bark [31]. The present method made no assumptions on the central phonetic processing of speech. The results reflect only the statistical correlation between adjacent frequency bands in the external speech signal, but the frequency range of the correlation can be measured more easily with other methods. It is interesting to speculate that the observed rather broad phonetic spectral integration may be influenced by these statistical characteristics of normal speech.

The difference between calculated high and low MI bounds is very small. Of course, the estimated bounds are still only approximate, as they are based on a Monte Carlo procedure. Test-retest differences were less than about 0.05 bit/s. Exact bounds were calculated only for conditioning sequences of length $D = 2$, because of the computational effort. For example, the exact low/high bounds were 7.70 and 13.58 bits/s at speech/noise ratio of 70/70 dB SPL, whereas the approximate bounds for $D = 15$ were 12.1 and 12.2 bits/s. Thus, the Monte Carlo approximation gives a much better estimate, even considering the random variability.

The present method relies on an automatic HMM training procedure to determine a set of "phonetic" classes in the clean speech signal. This procedure tends to assign speech frames to the same class, if they are spectrally similar and/or temporally adjacent. The obvious advantage is that the method can be applied to any speech signal and does not require laborious manual phonetic labelling of each signal frame. On the other hand, the automatic procedure may erroneously interpret some phonetically significant differences only as random variations within the same phonetic class. Conversely, it may assign information value to some phonetically insignificant variations. By omitting the first three computational steps in section 2.2.2, the method can easily be adapted to use manually labelled speech.

The proposed method for sensory information-rate estimation requires input speech and noise signals to be recorded on separate channels. The method can be applied to a wide variety of non-linear acoustic signal processing and psychophysical or physiological sensory models, with only the following restrictions: (1) The same temporal signal segmentation must be used for both the clean unprocessed speech and the sensory output. (2) The sensory model must specify covariance estimates for the internal sensory noise that limits signal discrimination performance. (3) The results should be interpreted with caution, if non-linear acoustic or sensory processing involves time-constants longer than the analysis frames. This may introduce systematic temporal correlations that will be represented as random variations in the hidden Markov model.

5. Conclusion

A calculation method was proposed which gives good approximate estimates of the rate of information (in bits/s) successfully transmitted from a speech source to the modelled neural output of the peripheral sensory system. This information rate sets definite upper limits on the listener's speech-recognition performance.

Calculations for a Swedish word-recognition test material, with a non-linear excitation-pattern auditory model, were consistent with speech recognition results obtained by normal-hearing listeners in speech-shaped masking noise. This suggests that the scarcity of sensory information may be the primary limiting factor in this test condition. Similar calculations for low-pass- and high-pass-filtered clean speech indicated a higher sensory information rate than required for the listeners' actual performance. These results suggest that the speech recognition performance under masking and filtering may be limited by different mechanisms.

Calculations were consistent with data in the literature, suggesting that information conveyed by adjacent narrow frequency bands is not additive, whereas widely separated low-pass and high-pass bands seem to convey almost completely additive information contributions.

References

[1] H. Fletcher: Speech and hearing in communication. Van Nostrand, New York, 1953.

[2] N. R. French, J. C. Steinberg: Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. **19** (1947) 90–119.

[3] ANSI-S3.5: American national standard methods for the calculation of the speech intelligibility index. American National Standards Institute, New York, 1997.

[4] T. Houtgast, H. J. M. Steeneken: The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acustica **28** (1973) 66–73.

[5] H. J. M. Steeneken, T. Houtgast: A physical method of measuring speech-transmission quality. J. Acoust. Soc. Am. **67** (1980) 318–326.

[6] T. Houtgast, H. J. M. Steeneken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. **77** (1985) 1069–1077.

[7] C. V. Pavlovic, G. A. Studebaker, R. L. Sherbecoe: An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. J. Acoust. Soc. Am. **80** (1986) 50–57.

[8] L. Magnusson: Predicting the speech recognition performance of elderly individuals with sensorineural hearing impairment. Scand. Audiol. **25** (1996) 215–222.

[9] D. Byrne, H. Dillon, T. Ching, R. Katsch, G. Keidser: NAL-NL1 procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures. J. Am. Acad. Audiol. **12** (2001) 37–51.

[10] L. Magnusson, M. Karlsson, A. Leijon: Predicted and measured speech recognition performance in noise with linear amplification. Ear Hear. **22** (2001) 46–57.

[11] I. J. Hirsch, E. G. Reynolds, M. Joseph: Intelligibility of different speech materials. J. Acoust. Soc. Am. **26** (1954) 530–538.

[12] H. J. M. Steeneken, T. Houtgast: Mutual dependence of octave-band weights in predicting speech intelligibility. Speech Communication **28** (1999) 109–123.

[13] H. Müsch, S. Buus: Using statistical decision theory to predict speech intelligibility. I. Model structure. J. Acoust. Soc. Am. **109** (2001) 2896–2909.

[14] H. Müsch, S. Buus: Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance. J. Acoust. Soc. Am. **109** (2001) 2910–2920.

[15] A. Leijon: Predicted speech intelligibility and loudness in model-based preliminary hearing-aid fitting. – In: Psychoacoustics, Speech and Hearing Aids. B. Kollmeier (ed.). World Scientific, Singapore, 1996, 123–132.

[16] A. Leijon: Model estimation of auditory-visual speech information transmission. European Conference on Audiology, Noordwijkerhout, The Netherlands, 1995, 238–242.

[17] T. M. Cover, J. A. Thomas: Elements of information theory. John Wiley and Sons, New York, 1991.

[18] B. Mandelbrot: An informational theory of the statistical structure of language. – In: Communication Theory. W. Jackson (ed.). Butterworths, London, 1953, 486–502.

[19] C. E. Shannon: Prediction and entropy of printed English. Bell Syst. Techn. J. **30** (1951) 50–64.

[20] G. A. Miller, G. A. Heise, W. Lichten: The intelligibility of speech as a function of the context of the test materials. J. Exp. Psychol. **41** (1951) 329–335.

[21] Y. Linde, A. Buzo, R. M. Gray: An algorithm for vector quantizer design. IEEE Trans. Commun. **COM-28** (1980) 84–95.

[22] L. R. Rabiner: A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE **77** (1989) 257–286.

[23] E. A. G. Shaw, M. M. Vaillancourt: Transformation of sound pressure level from the free field to the eardrum presented in numerical form. J. Acoust. Soc. Am. **78** (1985) 1120–1123.

[24] R. J. Baker, S. Rosen, A. M. Darling: An efficient characterisation of human auditory filtering across level and frequency that is also physiologically reasonable. – In: Psychophysical and physiological advances in hearing. 11th Int symposium on hearing. Whurr, Grantham, U.K., 1997, 81–88.

[25] S. Rosen, R. J. Baker, S. Kramer: Characterizing changes in auditory filter bandwidth as a function of level. – In: Auditory Physiology and Perception. Y. Cazals, K. Horner, L. Demany (eds.). Pergamon Press, Oxford, 1992, 429–446.

[26] B. C. J. Moore: Frequency analysis and masking. – In: Hearing. B. C. J. Moore (ed.). Academic Press, San Diego, 1995, 161–205.

[27] M. Florentine, S. Buus: An excitation-pattern model for intensity discrimination. J. Acoust. Soc. Am. **70** (1981) 1646–1654.

[28] A. J. M. Houtsma, N. I. Durlach, L. D. Braida: Intensity perception. XI. Experimental results on the relation of intensity resolution to loudness matching. J. Acoust. Soc. Am. **68** (1980) 807–813.

[29] B. Hagerman: Sentences for testing speech intelligibility in noise. Scand. Audiol. **11** (1982) 79–87.

[30] I. Holube, B. Kollmeier: Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. **100** (1996) 1703–1716.

[31] L. A. Chistovich: Central auditory processing of peripheral vowel spectra. J. Acoust. Soc. Am. **77** (1985) 789–805.