

FAST AMPLITUDE COMPRESSION IN HEARING AIDS IMPROVES AUDIBILITY BUT DEGRADES SPEECH INFORMATION TRANSMISSION

Arne Leijon and Svante Stadler

Sound and Image Processing Lab., School of Electrical Engineering, KTH
SE-10044, Stockholm, Sweden
phone: + (46) 8 790 7556, fax: + (46) 8 791 7654, email: arne.leijon@ee.kth.se
web: www.ee.kth.se

ABSTRACT

Common types of hearing impairment are caused mainly by a loss of nearly instantaneous compressive amplification in the inner ear. Therefore, it seems plausible that the loss might be compensated by fast frequency-dependent compression in the hearing aid.

We simulated impaired listeners' auditory analysis of hearing-aid processed speech in noise using a functional auditory model. Using hidden Markov signal models, we estimated the mutual information between the phonetic structure of clean speech and the neural output from the auditory model, with fast and slow versions of hearing-aid compression. The long-term speech spectrum of amplified sound was identical in both systems, as specified individually by the widely accepted NAL prescription for the gain frequency response.

The calculation showed clearly better speech-to-auditory information transmission with slow quasi-linear amplification than with fast hearing-aid compression, for speech in speech-shaped noise at signal-to-noise ratios ranging from -10 to $+20$ dB.

1. INTRODUCTION

The most common type of hearing impairment is characterized by a loss of the compressive amplification of cochlear outer hair cells [12]. This results in a reduced auditory dynamic range and abnormal loudness perception. The normal biological auditory compression acts nearly instantaneously [15]. Therefore, it has often been assumed that the loss of internal fast compression should be compensated by fast frequency-dependent amplitude compression in the external hearing aid [17, 8].

All modern hearing instruments are non-linear in the sense that they adapt their behaviour depending on the input signal, for example to reduce the audibility of background noise, separate speech from noise, sharpen speech formants, and/or emphasise weak consonants.

A hearing-aid compression system is usually called "fast" or "syllabic", if it adapts quickly enough to provide different gain frequency responses for adjacent speech sounds with different short-time spectra. Instruments with slow-acting automatic gain control keep their gain frequency response nearly constant in a given speech-plus-noise environment, and thus preserve the differences between short-time spectra in an ongoing speech signal. Hearing-aid compressors usually have frequency-dependent compression ratios, because the hearing loss varies with frequency. The compressive variations of the gain frequency response are

usually controlled by the input signal levels in several frequency bands.

Compression may be applied in hearing aids for several different purposes:

1. Present speech at comfortable level and spectral balance, compensating for variations in speaker distance and voice characteristics.
2. Protect the listener from sudden loud noises that would be uncomfortably loud if amplified with the gain frequency response needed for normal conversational speech.
3. Improve speech understanding by making also very weak speech segments audible, while still presenting louder speech segments at a comfortable level and spectral balance.

A fast compressor can to some extent meet all three objectives, whereas a slow compressor can only fulfil the first goal. A slow compressor must be supplemented with other systems to solve the other two problems.

Our present research question is focused only on the third objective: Assuming that the signal from a given individual speaker is already presented at a suitable level and with optimal spectral balance for a given hearing-impaired listener, and assuming also that there are no sudden loud sounds in the environment, should we then expect that a fast compressor can provide better speech understanding than a slow compressor?

Fast hearing-aid compression has two opposing effects with regard to speech recognition:

1. It provides additional amplification for weak speech components that might otherwise be inaudible.
2. It decreases spectral contrast between audible speech sounds.

It is not clear which of these two effects dominates in determining the net benefit for the user. Evaluation studies have generally failed to show a clear benefit of fast vs. slow-acting compression (for reviews, see [2, 6]).

In this study we estimate a theoretical measure of the benefit to be expected with fast compression. We simulate the listener's auditory analysis of hearing-aid processed speech in noise using a computational auditory model. The simulation allows us to estimate the mutual information (MI) between the phonetic structure of a sequence of spoken words and the corresponding stream of excitation patterns that mimics the neural output from the peripheral auditory system. The result indicates an upper bound on the speech-recognition performance that might be achieved by an ideal speech pattern classifier, trained on the auditory-model output.

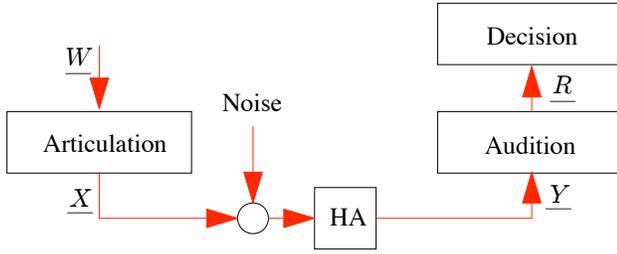


Figure 1: Model of the speech communication chain. The listener will try to estimate the word sequence \underline{W} from the stream \underline{R} of auditory neural patterns. A good hearing aid (HA) should increase the amount of speech information carried by the auditory data.

Of course, a given individual listener may not be able to process the sensory data optimally. One study indicates that some forms of compression are beneficial only for listeners with good cognitive functions [11]. Thus, if a theoretical simulation suggests that fast compression allows more auditory speech information to be conveyed to the brain, the slow (short-time linear) processing might still be better for some listeners. On the other hand, if the fast compressor provides *less* speech information to a theoretically optimal decoder, we may conclude that it probably cannot improve speech understanding for real listeners, compared to a similarly well-adapted slow compression system.

2. THEORY AND METHODS

Since neither speech nor audition is deterministic, it is natural to model each signal in the communication chain as a stochastic process. In our model, speech communication is summarized in the following stages (see Fig. 1):

1. The speaker decides to utter a word sequence \underline{w} , which is a realization of the random process \underline{W} .
2. The speaker pronounces the words, generating an acoustic signal, modeled by the random process \underline{X} .
3. The speech signal is contaminated by additive noise and processed by a hearing aid (HA).
4. The noisy processed signal \underline{Y} is analysed by the listener's peripheral hearing, producing an auditory pattern sequence \underline{r} , which is a realization of the random process \underline{R} .
5. The listener tries to estimate \underline{w} from the information available in \underline{r} .

We use this model structure to estimate the amount of speech information made available to hearing-impaired listeners using simulated hearing aids with fast and slow compression. Each of the stages in the chain is explained in further detail in the following subsections. A similar model structure was used in [16].

2.1 Speech Material

For all calculations we use a standardised Swedish closed-set speech recognition test material, usually called “Hagerman sentences” [7]. Similar tests exist in several other European languages. The test consists of 50 words, organised into 5 positions in each sentence, with ten possible words at each position. The words are chosen so that randomly selecting

one of the ten words at each position always generates a syntactically correct, but semantically unpredictable, sentence. We used this material mainly because the “grammar” defining all possible sentences can be described by a very simple probabilistic model. All calculations were performed for an acoustic speech presentation level of 65 dB re. 20 μ Pa.

The standard recording includes a separate channel with (slightly modulated) noise with a power density spectrum identical to the long-term spectrum of the test words. The speech and noise channels were mixed together at signal-to-noise ratios (SNR) ranging from -10 to $+30$ dB.

2.2 Signal Representation

The speech and noise waveforms are divided into 20-ms frames with 50% overlap, i.e. with 10-ms update interval. The power spectral density is computed for each frame at a non-linear frequency scale from 50 to 15000 Hz, with steps corresponding to 0.5 auditory equivalent rectangular bandwidth (ERB) [12]. Each short-time spectrum is converted to the dB domain, which is preferred here because Euclidean distance in this space is approximately proportional to perceptual difference. The spectrum is converted by an orthogonal discrete cosine transform (DCT) into 25 cepstral coefficients, because these are approximately uncorrelated for speech, noise and most other pseudo-stationary signals [4]. This enables us to assume diagonal covariance matrices when training the probabilistic models. After the model training, however, we use the inverse DCT to return to the spectral domain for describing the input signal to the auditory model, after hearing-aid processing.

We denote the sequence of acoustic short-time spectra for the clean speech as a discrete-time random vector sequence $\underline{X} = (X_1, \dots, X_t, \dots, X_T)$. Each X_t is a vector in which element X_{it} represents the spectrum level (in dB) sampled at frequency f_i and time t . The speech signal is mixed with noise, processed by a hearing instrument, and then analysed by the auditory model. The auditory-model response is denoted $\underline{R} = (R_1, \dots, R_t, \dots, R_T)$, using the same frame steps as in the acoustic spectral sequence. In each auditory vector R_t , element R_{jt} mimics the activity, at time t , in a set of nerve fibres originating from an auditory “place” z_j along the inner-ear basilar membrane. These fibres have maximal sensitivity to sound frequencies at their characteristic frequency, but the response is also influenced by adjacent input frequencies, as determined by the model frequency resolution, described in Sec. 2.5.

2.3 Rate of Mutual Information

We model the characteristics of random sequences \underline{X} and \underline{R} by hidden Markov models (HMM), as described in Sec. 2.4. The hidden state sequence in the clean-speech model is denoted $\underline{S} = (S_1, \dots, S_T)$. Each state represents a cluster of similar short-time spectra indicating one type of phonetic speech segment.

The *Rate of Mutual Information (MI)* specifies the amount of speech information successfully transmitted through the chain. The MI rate (in bits per frame) from the phonetic state sequence to neural output is defined as

$$\begin{aligned} r_{SR} &= \lim_{t \rightarrow \infty} h(R_t | R_1, \dots, R_{t-1}) - h(R_t | S_t) = \\ &= \lim_{t \rightarrow \infty} E \left[2 \log \frac{f(R_t | S_t)}{f(R_t | R_1, \dots, R_{t-1})} \right], \end{aligned} \quad (1)$$

where $h(\cdot)$ is the differential entropy function, defined using the expectation operator $E[\cdot]$ and the probability density $f(\cdot)$. As the differential entropy is a logarithmic measure of variability, the MI rate can also be seen as a measure of *Modulation Transfer*.

Using the trained and transformed HMM as a random source, we simply generate a long random state sequence (s_1, \dots, s_T) and corresponding output data sequences (r_1, \dots, r_T) of speech and noise processed by the hearing aid and the auditory system. We then estimate the speech-to-neural MI by replacing the expectation in Eq. (1) by stochastic integration as

$$\hat{r}_{SR} = \frac{1}{T} \sum_{t=d+1}^{d+T} 2 \log \frac{f(r_t | s_t)}{f(r_t | r_1, \dots, r_{t-1})}. \quad (2)$$

Here the denominator is conveniently calculated using the Forward Algorithm with the HMM. The standard deviation (SD) of the MI estimate was monitored during calculations. For each presented MI data point we used at least 50000 frames, or more if needed to make the SD of the estimate less than 0.01 bit/frame. The first $d = 20$ frames of each generated sequence were discarded in order to reach approximately stationary conditions in the averaged data.

2.4 Hidden Markov Model Training

Each word in the speech test material (Sec. 2.1) is modelled by a left-right HMM. We use tied HMM:s where all word models share the same set of output density functions. The conditional probability density for any observed K -dimensional vector r_t , from a HMM state $S_t = n$, is modelled as a Gaussian mixture (GMM)

$$f(r_t | S_t = n) = \sum_{m=1}^M w_{nm} \frac{1}{(2\pi)^{K/2} \sqrt{\det C_m}} e^{-\frac{1}{2}(r_t - \mu_m)^T C_m^{-1} (r_t - \mu_m)} \quad (3)$$

Note that the GMM component parameters μ_m and C_m are tied, i.e. identical for all states. Only the GMM weight factors w_{nm} depend on the HMM state. This HMM variant is sometimes called “semi-discrete”. The presented results were calculated with $M = 40$ GMM components, and each excitation-pattern vector r_t was calculated with $K = 75$ elements representing “place” samples with a uniform resolution of 0.5 ERB.

Each word HMM is first trained using the clean speech signal. This reduces the risk of over-fitting the HMM:s to the training data, as all GMM components are trained on the entire database. Then the noise signal is added at the specified SNR, the signal is processed by the simulated hearing aid, and the shared Gaussian components are retrained on the modified signal \underline{Y} . The auditory model defines a non-linear memoryless mapping of each input spectrum vector Y_t to a corresponding excitation-pattern vector $R_t = g(Y_t)$.

The model for the auditory pattern sequence is then obtained by modifying the shared Gaussian components to account for auditory-model transmission. This transformation is approximated by a locally linear expansion around the mean for each component, as

$$R_t \approx g(\mu_m) + D(Y_t - \mu_m) + W_t, \quad (4)$$

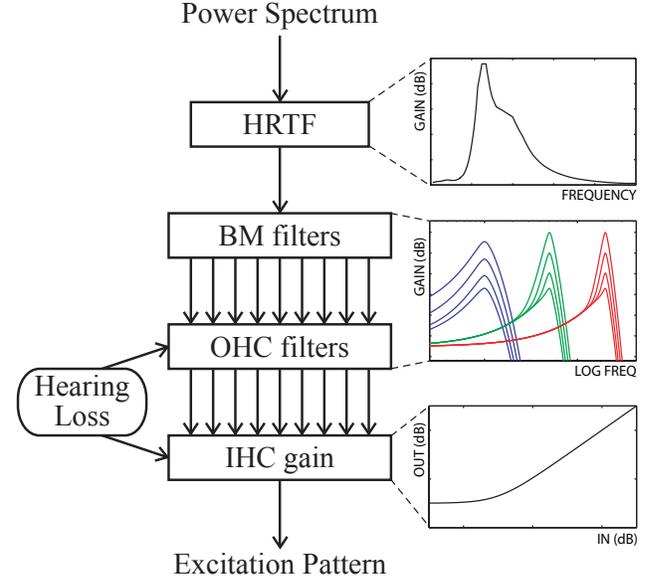


Figure 2: Structure of the auditory model. Typical transfer functions for a normal-hearing listener are plotted next to each stage. Sensorineural hearing loss is modelled mainly by reduced outer hair cell (OHC) non-linear gain, and if needed also by reduced sensitivity of inner hair cells (IHC).

where D is the partial-derivative matrix of the non-linear transformation $g(\cdot)$, and W_t is an additive noise vector that represents all neural random variations that are independent of the input signal.

All Markov state transition probabilities were kept unchanged in these transformations. After this training procedure, all word models are joined to form one single ergodic sentence HMM, including the transition probabilities between words in the probabilistic “grammar” for the speech material (Sec. 2.1).

2.5 Auditory Model

The auditory model captures essential features of peripheral hearing. It includes effects of outer ear transmission to the eardrum, middle-ear transmission, non-linear cochlear filtering, outer hair cell amplification, and inner hair cell sensitivity. The model operates entirely in the frequency domain, in a similar way as [13]. The structure of the model is summarised in Fig. 2. This model does not attempt to describe any retro-cochlear auditory processes, such as temporal integration, or any masking phenomena with time constants greater than about 20 ms. We only intend to estimate the sensory information available as input to central processes.

For each input short-time spectrum the model calculates a corresponding output excitation-level pattern. Cochlear filtering is modelled by a combination of two *Roex(p)*-shaped filters [1, 10], a linear tail part and a non-linear peak filter with gain depending on the peak-filter output. The peak filters are symmetric with normal ERB, and the bandwidth is independent of input level. This is a reasonable first approximation [1]. The tail filter has a fixed *roex* slope parameter $p = 8$ towards the low-frequency side. The maximum OHC gain was set to 50 dB at high frequencies, and

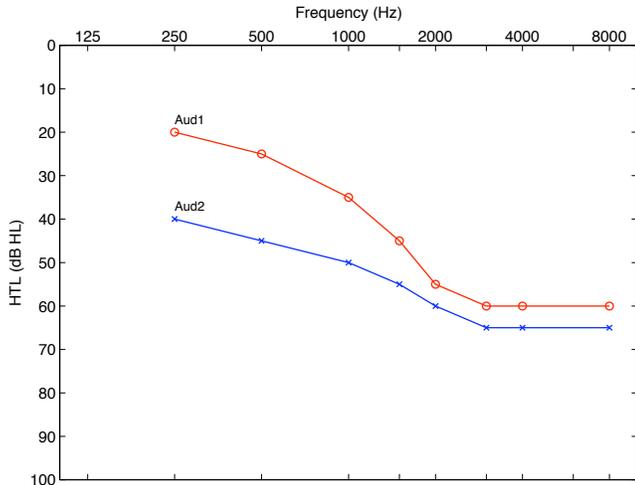


Figure 3: Audiogram with pure-tone hearing threshold losses for two simulated profiles of sensorineural hearing impairment. The normal threshold is 0 dB HL.

reduced at frequencies below 500 Hz as defined by Fig. 4 in [13]. Each short-time excitation pattern was evaluated at $K = 75$ cochlear “places” with best frequencies in the range 50 – 15000 Hz, with a resolution of 0.5 ERB.

We use the additive random noise W_i in Eq. (4) to represent the neural and perceptual variability that limits auditory spectral discrimination. These sensory noise components were statistically independent across both time and cochlear “place”, with a variance adjusted to reproduce normal intensity discrimination for broadband noise [10, 9].

2.6 Hearing Impairment

Auditory models were defined for two different types of hearing impairment, shown in Fig. 3. In most cases the only available data on hearing loss is the Hearing Threshold Level (HTL), which shows the pure-tone hearing threshold at a number of test frequencies. A threshold elevation can be caused by several different physiological changes, but in general these are not known individually. We have assumed that threshold elevations are caused primarily by a loss of normal outer hair cell gain for weak sounds. If the threshold loss exceeded the maximal OHC gain, the remainder of the loss was attributed to a loss of inner hair cell (IHC) sensitivity.

In our model, the loss of OHC gain automatically implies a loss of auditory frequency resolution, because the gain is reduced for the non-linear peak filter, and the remaining linear tail filter has a much wider response characteristic. No additional loss of frequency resolution was assumed.

2.7 Simulated Hearing-aid Processing

A hearing aid with *slow compression* was simulated simply by linear frequency-dependent amplification adapted to the hearing loss according to the NAL-R prescription rule [3], often used as a reference hearing-aid setting. This simulation is realistic, as any real slow compression system would adapt optimally to the overall level and long-term spectrum of the speaker, and we used a speech material with a single speaker at a fixed presentation level.

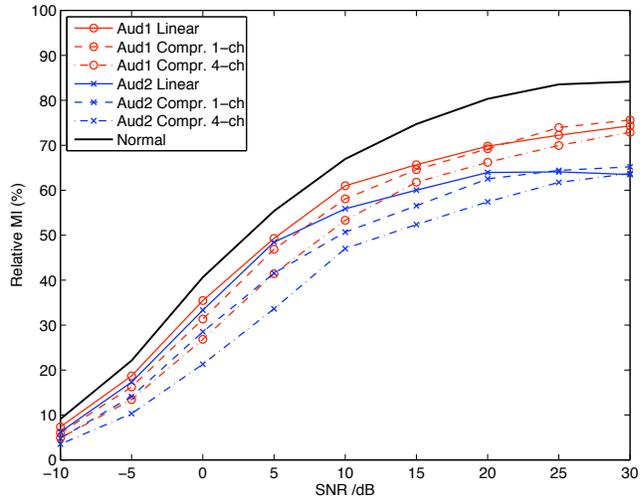


Figure 4: Relative amount of received auditory speech information as a function of signal-to-noise ratio (SNR). The mutual information (MI) is calculated between clean speech HMM states and the auditory response to noisy speech, processed by hearing aids with fast compression and with linear amplification, adapted to two types of hearing loss. The MI is shown in percent of the state entropy for clean speech.

A hearing aid with *fast compression* was simulated by adapting the gain frequency response depending on the short-time spectrum in each 10-ms time frame (20 ms duration incl. overlap), independently of previous frames. However, it would be unrealistic to adapt the gain independently at each frequency sample, because this would reduce fine spectral details. Instead, we controlled the adaptive gain by a smoothed version of the spectrum, estimated by including only the lowest-order cepstrum coefficients. We used either 1 or 4 coefficients to simulate generic forms of frequency-dependent compression with either a single channel or 4 channels. Compression Ratios (CR) were set depending on the hearing threshold loss in dB (HTL) at each frequency, as $CR = 100/(100 - HTL)$, but limited to $1 \leq CR \leq 2.5$. The compression ratios were independent of input levels, because the auditory model also assumed level-independent CR. Finally the overall linear gain frequency response was adjusted to reproduce the compressed signal with exactly the same long-term power spectrum as for the slow compression.

3. RESULTS AND DISCUSSION

The estimated rate of mutual information (MI) between clean speech segment states (phonetic categories) and peripheral auditory response to noisy hearing-aid-processed speech is shown in Fig. 4 for the types of hearing loss shown in Fig. 3. Here the MI results are plotted as a percentage of the source state entropy which was 0.487 bits/frame or 48.7 bits/s. The MI is clearly lower with fast compression than with the linear amplification simulating slow compression, for all SNR:s less than 20 dB.

The calculated results in Fig. 4 indicate that the use of fast compression destroys more speech information than the amount of information gained by increasing the audibility of weak speech segments. This result might have been antici-

pated at low SNR values, where speech audibility is limited mainly by noise masking rather than by the hearing threshold. However, the reduction of speech information is apparent also at rather high SNRs.

The result should be interpreted with some caution, because we simulated a slow compression system by linear amplification, adapted to the given speech presentation level in an acoustic environment with only speech-shaped steady noise and with no loud transient sounds. There are methods to combine slow compression for speech with fast transient reduction, e.g. [14], but many practical hearing-aid implementations of slow-acting compression use a long release time and short attack time, because the compressor is also intended to protect the listener from sudden loud noises. Such a system reduces the amplification of speech throughout the release period after a loud sound, and may therefore lose more information in the weak speech components than indicated by our simulation.

The model preferred in [5] might account better for a loss of temporal resolution but does not include the effect of hearing loss on frequency resolution. Our results may not be applicable for a listener with impaired peripheral temporal resolution. However, as the fast compressor degraded speech information, assuming an ideal central detector, it seems unlikely that it can improve speech recognition for listeners with impaired central auditory functions.

4. CONCLUSION

We calculated the amount of speech information successfully transmitted through a functional model of the peripheral auditory system of hearing-impaired listeners using hearing aids with either fast compression or slow quasi-linear amplification, adapted to the individual hearing loss.

Although the hearing impairment was modelled mainly as a loss of the normal fast compression in the inner ear, the calculation showed clearly better speech-to-auditory information transmission with linear amplification than with fast hearing-aid compression, for speech in speech-shaped noise at signal-to-noise ratios ranging from -10 to $+20$ dB.

REFERENCES

- [1] R. J. Baker and S. Rosen, "Auditory filter nonlinearity across frequency using simultaneous notched-noise masking," *Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 454–462, 2006.
- [2] L. Braida, N. Durlach, R. Lippman, B. Hicks, W. Rabinowitz, and C. Reed, *Hearing Aids - a Review of Past Research on Linear Amplification, Amplitude Compression and Frequency Lowering (Monograph no 19)*. Rockville, MD: American Speech and Hearing Association, 1979.
- [3] D. Byrne and H. Dillon, "The national acoustic laboratories (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing*, vol. 7, pp. 257–265, 1986.
- [4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.
- [5] R. P. Derleth, T. Dau, and B. Kollmeier, "Modeling temporal and compressive properties of the normal and impaired auditory system," *Hearing Research*, vol. 159, pp. 132–149, 2001.
- [6] A. Goedegebure, "Phoneme compression. Processing of the speech signal and effects on speech intelligibility in hearing-impaired listeners," Ph.D. dissertation, Erasmus Univ Rotterdam, 2005.
- [7] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian Audiology*, vol. 11, pp. 79–87, 1982.
- [8] T. Herzke and V. Hohmann, "Effects of instantaneous multi-band dynamic compression on speech intelligibility," *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 3034–3043, 2005.
- [9] A. Houtsma, N. Durlach, and L. Braida, "Intensity perception. XI. Experimental results on the relation of intensity resolution to loudness matching," *Journal of the Acoustical Society of America*, vol. 68, no. 3, pp. 807–813, 1980.
- [10] A. Leijon, "Estimation of auditory information transmission capacity using a hidden Markov model of speech stimuli," *Acustica - Acta Acustica*, vol. 88, no. 3, pp. 423–432, 2002.
- [11] T. Lunner and E. Sundevall-Thorén, "Interactions between cognition, compression, and listening conditions: Effects on speech-in-noise performance in a two-channel hearing aid," *Journal of the American Academy of Audiology*, vol. 18, no. 7, pp. 604–617, 2007.
- [12] B. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London: Academic Press, 2003.
- [13] B. C. Moore and B. R. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss," *Hearing Research*, vol. 188, pp. 70–88, 2004.
- [14] P. Nordqvist and A. Leijon, "Hearing-aid automatic gain control adapting to two sound sources in the environment, using three time constants," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3152–3155, 2004.
- [15] A. Recio, N. C. Rich, S. S. Narayan, and M. A. Ruggero, "Basilar-membrane responses to clicks at the base of the chinchilla cochlea," *Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 1972–1989, 1998.
- [16] S. Stadler, A. Leijon, and B. Hagerman, "An information theoretic approach to estimate speech intelligibility for normal and impaired hearing (poster nr. 10)," in *Interspeech 07*, Antwerpen, BE, 2007. [Online]. Available: <http://www.interspeech2007.org>
- [17] E. Villchur, "Signal processing to improve speech intelligibility in perceptive deafness," *Journal of the Acoustical Society of America*, vol. 53, pp. 1646–1657, 1973.