An Information Theoretic Approach to Predict Speech Intelligibility for Listeners with Normal and Impaired Hearing

Svante Stadler[1], *Arne Leijon*[1], *Björn Hagerman*[2]

[1]Sound and Image Processing Lab, Royal Institute of Technology,Stockholm, Sweden
[2]Unit of Technical and Clinical Audiology, Karolinska Institutet, Danderyd, Sweden
{svante.stadler, arne.leijon} @ee.kth.se, bjorn.hagerman@cns.ki.se

10.21437/Interspeech.2007-197

Abstract

A computational method to predict speech intelligibility in noisy environments has been developed. By modeling speech and noise as stochastic signals, the information transmission through a given auditory model can be estimated. Rate-distortion theory is then applied to predict speech recognition performance. Results are compared with subjective tests on normal and hearing impaired listeners. It is found that the method underestimates the supra-threshold deficits of hearing impairment, which is believed to be due to an overly simple auditory model and a small dictionary size.

Index Terms: speech intelligibility, speech perception, auditory models

1. Introduction

The Speech Intelligibility Index (SII) is a standardized method for predicting speech intelligibility in noisy environments [1]. It is popular due to its simplicity and its accuracy of prediction for a wide range of noise types, but it also has a number of inherent drawbacks:

- Stationary noise is assumed.
- It is based on empirical results; the underlying principles are not known.
- It cannot account for any nonlinear signal modifications (e.g. noise reduction and dynamic compression).
- It does not account for the supra-threshold deficits of hearing impairment.

The first issue was partially resolved by Rhebergen et al [2], by calculating the SII for every block of speech, and averaging over the entire signal. The results agree well with subjective measurements on normal hearing listeners, but it has not been validated for hearing impaired listeners.

Several researchers have proposed corrections to the SII that account for supra-threshold deficits. This correction term is usually called the "desensitization" factor, because cochlear hearing loss seems to lead to reduced sensitivity to spectral differences, although the sound level is above the threshold of hearing at all frequencies. These methods improves prediction, but they are generally empirical solutions which do not claim to explain the physiological cause of the phenomenon. (In the following results, we use the method described in [3] as a reference.)

The proposed method aims to overcome all of these restrictions by viewing human speech communication as any communication channel, where the information transmission constraints found by Shannon apply. By using a model of the source (speech), the channel (noisy environment and the peripheral auditory system) and the receiver (the human brain), we can apply information theory and rate-distortion theory to predict speech intelligibility. This approach was first introduced in [4], and this paper is a continuation and refinement of that work.

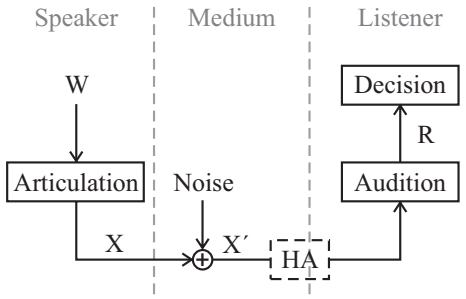


Figure 1: Model of Speech Communication. The listener will try to estimate the word sequence W from the auditory sequence R. A potential hearing aid (HA) will hopefully improve the estimate.

The goal of this study is to construct a model of human speech communication that can predict speech intelligibility in noisy environments, given only a model of the listener's auditory system; that is, *without* adaptation to experimental speech recognition data. The method will be evaluated by comparing results with experimental data.

2. Model of Speech Communication

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

1. The talker decides to utter the word sequence w, which is a realization of the random process W.
2. The talker pronounces the words, generating an acoustic signal, modeled by the random process X.
3. The speech signal is contaminated by additive noise N.
4. The signal reaching the listener's ear is $X' = X + N$.
5. The noisy signal passes through the listener's periphal hearing, producing an auditory pattern sequence r, which is a realization of the random process R.
6. The listener tries to estimate w from the information available in r.

With this structure, it is simple to insert an extra processing stage between 4 and 5, which would represent a hearing aid or

some other form of signal processing. Each of the stages in the chain is explained in further detail in the following subsections.

2.1. Signal Representation

The signal waveform is divided into 20 ms, 50% overlapping frames. The power spectral density is computed in 0.5 auditory equivalent rectangular bandwidth (ERB) steps from 50 to 15000 Hz and converted to dB. This domain is preferred, because Euclidean distance in this space is approximately proportional to perceptual difference [5]. From this we compute 25 cepstral coefficients, which are approximately uncorrelated for speech, noise and most other pseudo-stationary signals [6]. This representation enables us to assume diagonal covariance matrices, which is crucial when the training data set is small. After the model training, however, we use the inverse DCT to return to the spectral domain.

It is a basic assumption of this model that the ERB-scale spectrum carries all the information that is relevant for speech perception. This is not entirely true, as pitch and temporal fine structure carry speech information as well [7]. However, it is not clear whether these cues are useful in very noisy conditions or not.

2.2. Probabilistic Speech Model

It is important that we do not use a deterministic speech signal, because the listener does not know exactly how each word is pronounced by the talker. Therefore, we aim to make the model account for natural variations in pronunciation. Since we in general only have access to one particular utterance of each word, we need a model that is inexact by nature. To this end, we use a Hidden Markov Model (HMM) with a *tied* Gaussian Mixture Model (GMM), where each word is represented by a left-right HMM, and all HMMs share a GMM for their output distributions; only the mixture weights are adapted for each word. This reduces the risk of over-fitting the HMM's to the training data, as the GMM components are trained on the entire database. Thereafter, all the word models are joined to one big sentence HMM. This is important because the listener does not know exactly where one word ends and the next begins, and this uncertainty should be built into the model.

Each word was resynthesized from their models, to verify that the models had caught the speech information with a reasonable amount of random variation. An informal listening test found that 50 GMM components and 20 Markov states per word is appropriate to make all words intelligible, but with random variations in phoneme spectrum and duration.

2.3. Model of Noisy Speech

Similarly to the speech model, an HMM is trained on the pure noise signal. Then we simply merge the speech and noise models into one HMM, to get a model of noisy speech. The models are expressed in the dB domain, but the sum must be performed in the power domain. This is a problem, because there is no closed-form expression for their sum. However, a good estimate is calculated by using a Monte Carlo method: random observations are generated from each of the distributions, then converted to power spectrum, added, and converted back to dB. Since the obtained vectors are observations from the desired distribution, we can simply calculate the mean and covariance from these vectors.

2.4. Auditory Model

In principle, the proposed method can make use of any auditory model, as long as it is possible to define the detectability, d', between two feature sets. For these measurements, we have implemented the auditory filter model proposed by Baker and Rosen [8], because it has been finely tuned to account for the nonlinear effects of cochlear hearing loss. The filtered signals are then passed through a simple inner hair cell (IHC) stage, where a threshold excitation is added to simulate spontaneous activity. The output is an "excitation pattern", which represents the neural activity at each place in the cochlea. Following the discrimination model proposed by Florentine and Buus in [9], the excitation pattern is finally converted to dB. In this space, the detectability is simply proportional to Euclidean distance. It has been verified that the model matches the experimental discrimination results for hearing impaired listeners from [10]. The structure of the model is summarized in Figure 2.

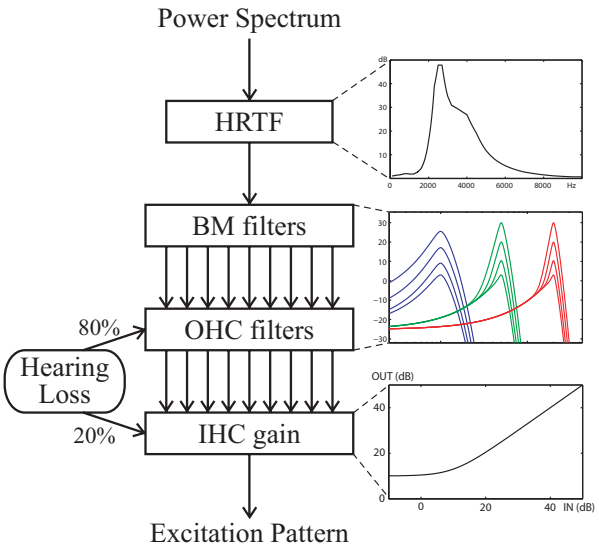


Figure 2: Structure of the Auditory Model. Typical transfer functions are plotted next to each stage.

Perhaps the greatest difficulty in setting up the auditory model is deciding on how to account for a specific hearing loss. In most cases, the only available data on hearing loss is the Hearing Threshold Level (HTL), which is the pure-tone hearing threshold at a number of frequencies. A threshold elevation can be attributed to a number of physiological damages, but in general these are not known on the individual level. Therefore, we make a somewhat educated guess and assume that 80% of the elevation is due to outer hair cell (OHC) degradation and 20% is due to IHC degradation. If the threshold exceeds the maximum gain of the OHC, the remainder of the loss is attributed to the IHC.

The model for the auditory pattern sequence R is obtained by taking the model for the noisy speech X' and modifying the output GMM by linearizing the auditory model around each component of the GMM. In effect, the auditory model becomes a piece-wise linear function of the power spectrum. It is assumed that this is enough to capture the nonlinearities of the auditory model, at least for a single signal-to-noise ratio (SNR) and overall presentation level.

2.5. Estimation of Information Transmission

To quantify the information transmission capability of a channel, it is common the to use Mutual Information (MI) (see, e.g., [11]). In this case, the objective is to estimate the MI between the word sequence W and the corresponding auditory excitation pattern R. This MI can be expressed as

$$I(W;R) = E\left[\log\left(\frac{f_{R|W}(r|w)}{f_R(r)}\right)\right] \quad (1)$$

where f represents a probability density and E represents the expectation value over W and R. There is no closed-form solution for this expression, but it can be estimated with arbitrary precision using stochastic integration. To this end, we simply generate random word sequences w_i, which is used to generate an acoustic sequence x_i, which is mixed with noise and passed through the auditory model, which gives us the excitation pattern sequence r_i. Now our estimate simply becomes

$$I(W;R) \approx \frac{1}{N}\sum_{i=1}^{N} \log\left(f_{R|W}(r_i|w_i)\right) - \log\left(f_R(r_i)\right) \quad (2)$$

In the following results, N is set so that the 95% confidence interval is below 0.05 bits.

Once we have estimated the MI, we can use rate-distortion theory to find an upper bound on the recognition score. In this case, the rate is the MI, expressed in bits/word, and the distortion is the probability of error in the word sequence estimation. To calculate this relation, we must first assume a prior probability distribution for W, which should reflect the listener's expectation of what words she might hear. In a listening test such as the one described in the following section, it is reasonable to assume a uniform prior distribution, if the listener has been trained enough to learn all the alternatives. For natural sentences, however, it is more realistic to assume a "Zipf" distribution, which is the approximate frequency distribution of words in natural language [12]. These cases will give different rate-distortion functions, shown in Figure 3. The rate-distortion relations are found numerically using the Blahut algorithm [11]. They are upper bounds, only reachable if the listener has perfect memory and the auditory sequence is infinitely long. Although this is never true for human listeners, it can be a good approximation in many cases.

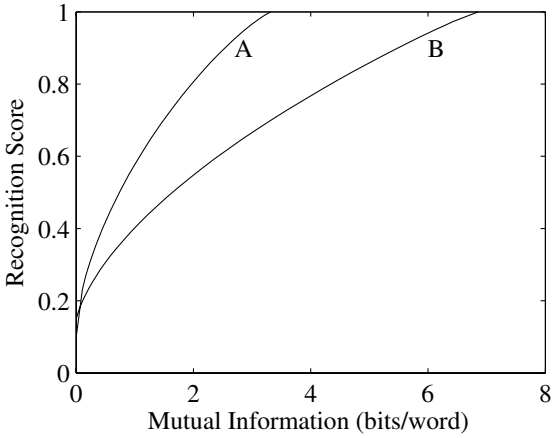


Figure 3: Rate-distortion function for: (A) a uniform distribution with ten words, and (B) a "Zipf" distribution with 500 words.

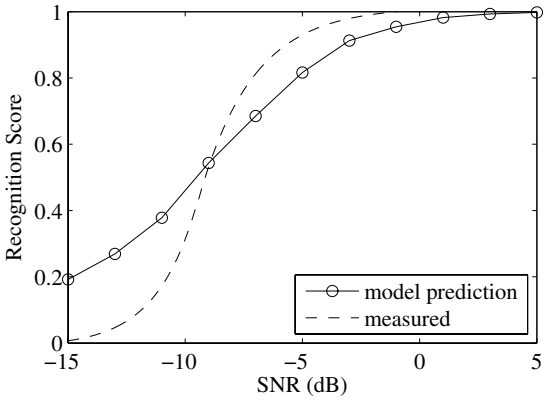


Figure 4: Average speech recognition score for normal hearing listeners (adapted from [13]) and model prediction, as a function of SNR.

3. Comparison with Experimental Data

The Hagerman's Sentences is a Swedish word recognition test which has been used to determine speech recognition thresholds in noise. Similar tests exist in several languages. The experimental data for normal hearing listeners are taken from [13] and the data for hearing impaired listeners are from [14].

The test consists of 50 words, organized into 5 positions with ten possible words for each position, so that choosing one of the ten words of each position generates a grammatically correct sentence. In [13], sentences were mixed with slightly modulated, speech-weighted noise and the subjects' recognition scores were noted for a number of SNR. In Figure 4, the experimental results are plotted along with model predictions. The underestimation at high SNR implies that too much information was discarded when constructing the speech model from the speech signal. The overestimation at low SNR gives support to the notion that human listeners are not capable of integrating small amounts of information.

In [14], 40 hearing impaired subjects were tested for speech recognition in slightly and fully modulated, speech-weighted noise. An adaptive procedure was employed to find the SNR level that corresponded to 2 out of 5 words correct, i.e. a recognition score of 0.4. In these measurements, the presentation level was set relative to the HTL so that the speech and noise were audible but not uncomfortable for all subjects. In Figure 5, predictions by the proposed model are plotted along with predictions by SII with and without desensitization. Only the data for slightly modulated noise has been used here. The proposed method and the SII are comparable in performance, both having only a slight correlation with the experimentally determined thresholds. The SII with desensitization is more accurate; although there is a large individual variance, the overall tendency is correct.

4. Discussion and Future Work

It is clear that this model, although quite accurate for normal hearing listeners, can only account for a fraction of the difference in speech recognition ability between normal and hearing impaired listeners. Many researchers suggest that this difference is mainly due to reduced cognitive ability, and the results

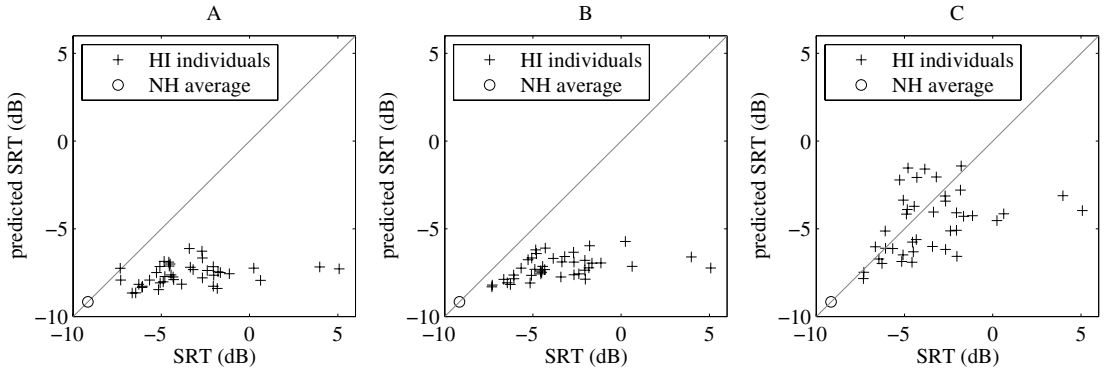


Figure 5: Speech recognition thresholds in noise (SRT) for 40 hearing impaired subjects (HI), average threshold for normal hearing subjects (NH), and model predictions. **A** Prediction by the proposed model. **B** Prediction by SII. **C** Prediction by SII with desensitization for hearing loss. To make a fair comparison, the respective MI and SII thresholds were set to match the normal hearing average.

presented here do not contradict them. The reduced frequency resolution caused by loss of OHC, which is generally believed to play a big part in this effect, can only account for a fraction of the difference according to these results. However, it is possible that the effect would be more dramatic with a larger speech database, because then the ability to detect subtle spectral differences would be more critical.

Prediction might also improve with a more detailed auditory model. One obvious drawback of the implemented model is that it does not include pitch perception, which can be used to discriminate between voiced and unvoiced phonemes. It is also possible that the periodicity helps the listener separate speech from background noise. This is an area which is in need of further research.

There are many auditory models available that can represent temporal fine structure, but few have been validated for both normal hearing and hearing impaired listeners, which is key in this type of application.

5. Conclusion

A model of human speech communication has been implemented, which tries to predict speech intelligibility in noisy environments. It produces results similar to SII, without any adaptation to experimental speech recognition data. It was found that the assumed causes of hearing impairment were not enough to explain the supra-threshold deficits of cochlear hearing impairment. Further research is needed to find the physiological cause of these deficits.

6. References

[1] ANSI-S3.5, *American national standard methods for the calculation of the speech intelligibility index*. New York: American National Standards Institute, 1997. [Online]. Available: http://www.sii.to/index.html

[2] K. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[3] C. Pavlovic, G. Studebaker, and R. Sherbecoe, "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," *Journal of the Acoustical Society of America*, vol. 80, pp. 50–57, 1986.

[4] 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.

[5] R. Plomp, *Aspects of Tone Sensation*. London: Academic Press, 1976.

[6] S. B. 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. ASSP-28, no. 4, pp. 357–366, 1980.

[7] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Philosophical Transactions: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[8] 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.

[9] M. Florentine and S. Buus, "An excitation-pattern model for intensity discrimination," vol. 70, no. 6, pp. 1646–1654, 1981.

[10] M. Florentine, C. Reed, W. Rabinowitz, L. Braida, N. Durlach, and S. Buus, "Intensity perception. xiv. intensity discrimination in listeners with sensorineural hearing loss," *Journal of the Acoustical Society of America*, vol. 94, pp. 2575–2586, 1993.

[11] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

[12] W. Li, "Random texts exhibit Zipf's law-like word frequency distribution," *IEEE Transactions on Information Theory*, vol. 38, 1992.

[13] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian Audiology*, vol. 11, pp. 79–87, 1982.

[14] ——, "Speech recognition threshold in slightly and fully modulated noise for hearing-impaired subjects," *International Journal of Audiology*, vol. 41, no. 6, pp. 321–329, 2002.