A Simple Model of Speech Communication and its Application to Intelligibility Enhancement

W. Bastiaan Kleijn, *Fellow, IEEE*, and R. C. Hendriks, *Member, IEEE*

***Abstract*—We introduce a model of communication that includes noise inherent in the message production process as well as noise inherent in the message interpretation process. The production and interpretation noise processes have a fixed signal-to-noise ratio. The resulting system is a simple but effective model of human communication. The model naturally leads to a method to enhance the intelligibility of speech rendered in a noisy environment. State-of-the-art experimental results confirm the practical value of the model.**

***Index Terms*—Enhancement, intelligibility, speech.**

I. Introduction

Modern communication technology allows a user to communicate from almost anywhere to almost anywhere. As the physical environment of the talker and the listener is not controlled, noise often affects the ability of the parties to communicate. We can distinguish two separate problems. On the one hand, the signal recorded by the microphone can be noisy. A large research effort has been dedicated to reducing the noise in the recorded signal either at the transmitter, e.g., [1]–[3], or at the receiver [4]. On the other hand, the sound is played back for the listener in a noisy environment. In recent years, a significant effort has been made towards improving the intelligibility of the sound played back in a noisy environment, e.g., [5]–[11]. We introduce a new paradigm for improving the intelligibility of speech played out in noisy environments.

The main innovations in this contribution are that *i*) we consider noise inherent in the message production process as well as noise inherent in the message interpretation process, *ii*) we consider the case where such inherent noise has a fixed signal-to-noise ratio. When production and interpretation noise are considered, information theory can be used to define a simple but effective model of human communication. This can then be used to design a state-of-the-art algorithm to optimize the intelligibility of speech in a noisy environment.

Production noise is typical of biological communication systems. For human communications, this can be seen at various levels of abstraction. The word choice to convey a message varies between occasions and talkers. At a lower level of abstraction, speech can be seen as a sequence of discrete set of phonemes and the pronunciation of these phonemes varies significantly from one utterance to the next. This variation is reflected in the fact that speech recognition uses statistical acoustic models, e.g., [12], [13]. The interpretation process for speech is also noisy: speech signals that are ambiguous in their pronunciation may be interpreted in various ways.

Information theoretical concepts have been used in the analysis of human hearing [14] and for the definition of measures of intelligibility [15]. These models do not have the notion of production noise, but the model of [14] considers sensory noise, which corresponds to our interpretation noise. The models of [14] and [15] appear not to have been used for optimizing intelligibility.

II. Model of the Communication Chain

We consider the transmission of a message S that is represented by a K-dimensional stationary discrete-time random process. The process is composed of real or complex scalar variables $S_{k,i}$, where $k \in \kappa$ is the dimension index and $i \in \mathbb{Z}$ is the time index. In the context of speech specified as a sequence of speech spectra, the variables $S_{k,i}$, may describe the complex amplitude or the gain in a particular time-frequency bin.

A. Model with Production and Interpretation Noise

Let the message have a "production" noise, representing the natural variation in its generation, either for a single person or across all talkers. The transmitted signal for dimension k at time i is then

$$X_{k,i} = S_{k,i} + V_{k,i}, \tag{1}$$

where $V_{k,i}$ is production noise. The received signals satisfy

$$Y_{k,i} = X_{k,i} + N_{k,i} \tag{2}$$

where $N_{k,i}$ is environmental noise. Finally, the received symbols are interpreted, which is also a noisy operation:

$$Z_{k,i} = Y_{k,i} + W_{k,i}, \tag{3}$$

where $W_{k,i}$ is "interpretation" noise. Note that $S \to X \to Y \to Z$ is a Markov chain.

The mutual information rate between the original multi-dimensional message sequence S and the received multi-dimensional message sequence Z describes the effectiveness of the communication process. In this first description, we assume the processes to be memoryless, which is reasonable for

Manuscript received April 21, 2014; revised June 26, 2014; accepted August 13, 2014. Date of publication August 28, 2014; date of current version September 29, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ali Taylan Cemgil

W. B. Kleijn is with Victoria University of Wellington, Wellington, New Zealand and also with the Delft University of Technology, Delft, The Netherlands (e-mail: bastiaan.kleijn@ecs.vuw.ac.nz).

R. C. Hendriks is with Delft University of Technology, Delft, The Netherlands (e-mail: r.c.hendriks@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2351784

time-frequency signal representations. The mutual information rate is then equal to the mutual information $I(S_i;Z_i)$ between the multi-dimensional symbols S_i and Z_i at a particular time instant i. We furthermore assume that the individual component signals of the multi-dimensional sequence are independent. Then we can write

$$I(S_i;Z_i) = \sum_{k\in\kappa} I(S_{k,i};Z_{k,i}). \tag{4}$$

Let us consider the behavior of the production and interpretation noises for the speech application. Speech production is a probabilistic process. A speech sound is never exactly the same. This variability is largely independent of the power level at which it is produced. That is, the production SNR $\frac{\sigma^2_{S_k}}{\sigma^2_{V_k}}$ is constant (with $\sigma^2_{S_k} = \mathrm{E}[S^2_k]$, where E denotes expectation and where we omit the time subscript i to simplify notation). It follows that the correlation coefficient between the message signal $S_{i,k}$ and the actual signal $X_{i,k}$, denoted as $\rho_{S_kX_k}$, is a fixed number on [0,1].

A fixed SNR for the interpretation noise is also reasonable. The auditory system contains a gain adaptation for each critical band [16], which means that the precision of the interpretation scales with the signal over a significant dynamic range. Thus, the interpretation SNR $\frac{\sigma^2_{Yk}}{\sigma^2_{W_k}}$ and the correlation coefficient $\rho_{Y_kZ_k}$ can be modeled as fixed.

The constant-SNR production and/or interpretation noise has a significant effect on a power constrained communication system. In a conventional communication system with parallel channels (without production and/or interpretation noise) the best information throughput is obtained by *waterfilling*[17]: more signal power is provided to communication channels with low noise power. However, in the present communication system there is generally little benefit to having a channel SNR, $\frac{\sigma^2_{X_k}}{\sigma^2_{N_k}}$, that is significantly beyond the production SNR or the interpretation SNR. The usefulness of a particular communication channel "saturates" near the production SNR or the interpretation SNR, whichever is lower.

When the new communication model is applied to speech, we must consider the particularities of the human auditory system. We distinguish the *acoustic* and *auditory* representation of the signal. The mapping $\mathcal{A}$ from the acoustic to the auditory representation is surjective. The frequency resolution of both the speech features and the auditory system varies with frequency. A typical scale is the ERB (equivalent rectangular bandwidth) scale, e.g., [18], [19]. It is natural, e.g., [15], to consider the auditory-domain signal to have one independent component signal per ERB. Auditory models provide a manner of deriving such component signals. We assume conceptually that the component signals associated with the ERB bands all have identical bandwidth. For example, we can reduce the ERB bands to one characteristic bandwidth ω_0 (which can remain unspecified in our application) by frequency translating bands of bandwidth ω_0, within an ERB band to the baseband and summing or integrating them. If we assume that the component frequencies of the original signal are independent, the component signal of bandwidth ω_0 representing an ERB band retains the power of the original signal within that ERB band.

B. Tractable Model that Includes Enhancement

We now insert a machine-based enhancement operator $\mathcal{G}$ in the Markov chain. If we mark by $\tilde{}$ all signals affected by the enhancement operator we get a Markov chain $S \rightarrow X \rightarrow \tilde{X} \rightarrow \tilde{Y} \rightarrow \tilde{Z}$, where $\tilde{X} = \mathcal{G}(X)$.

To formulate a tractable optimization problem, let us make the assumption that all processes are jointly Gaussian, stationary, and memoryless. For ease of notation, we omit the time index i from here-on forward. For the Gaussian case it can be shown that

$$I(S_k;\tilde{Z}_k) = -\frac{1}{2}\log(1-\rho^2_{S_k\tilde{Z}_k}). \tag{5}$$

We can make several simplifications. Exploiting the Markov chain property, we see that $\rho_{S_k\tilde{Z}_k} = \rho_{S_k\tilde{X}_k}\rho_{\tilde{X}_k\tilde{Y}_k}\rho_{\tilde{Y}_k\tilde{Z}_k}$. The fixed interpretation SNR implies $\rho_{\tilde{Y}_k\tilde{Z}_k} = \rho_{Y_kZ_k}$. If the enhancement operator $\mathcal{G}$ is an affine function for each component signal, then we also have $\rho_{S_k\tilde{X}_k} = \rho_{S_kX_k}$.

Next, we consider how the theory is affected if the signal is interpreted in its auditory representation. In Section II-A we described a mapping $\mathcal{A}$ from the acoustic to the auditory representation. Within each ERB band a number of Gaussian variables are combined into a single process. Our model without enhancement within a particular ERB band with index m consists of *i*) the generation of a set of variables S_k, $k\in\kappa_m$, *ii*) the addition of independent noise variables $U_k = V_k + N_k + W_k$ to each generated variable, and *iii*) the summation (in the ear) of all variables to the single ERB band random variable: $Z_m = \sum_{k\in\kappa_m} S_k + U_k$. Assuming $\rho^2_{S_k,S_k+U_k}$ is constant for $k\in\kappa_m$, it can then be shown that

$$I(\{S_k\}_{k\in\kappa_m};Z_m) = -\frac{1}{2}\log(1-\rho^2_{S_n,S_n+U_n}),\ n\in\kappa_m. \tag{6}$$

which is similar to (5) before the enhancement operator is added. Thus, we have found that under the forementioned assumptions the above theory carries over to the case where the final receiver is the human auditory system, which integrates within signal bands.

C. Relation to Classical Measures of Intelligibility

The measure (4) is related to existing heuristically-derived measures. If we write the channel SNR a $\xi_k = \frac{\sigma^2_{\tilde{X}_k}}{\sigma^2_{N_k}}$ and $\rho_{0,k} = \rho_{S_k,X_k}\rho_{Y_k,Z_k}$ we can use (5) to rewrite (4) as

$$I(S;\tilde{Z}) = -\sum_{k\in\kappa}\frac{1}{2}\log\left(\frac{(1-\rho^2_{0,k})\xi_k+1}{\xi_k+1}\right). \tag{7}$$

Using $I_k = -\frac{1}{2}\log(1-\rho^2_{0,k})$ and the sigmoid $A_k(\xi_k) = \frac{\log\frac{(1-\rho^2_{0,k})\xi_k+1}{\xi_k+1}}{\log(1-\rho^2_{0,k})}$ we obtain

$$I(S;\tilde{Z}) = \sum_{k\in\kappa} I_k A_k(\xi_k). \tag{8}$$

If we identify I_k as the scaled *band-importance function* and $A_k(\cdot)$ as the *weighting function* the mutual information can be interpreted as the scaled articulation index (AI), e.g., [20], [21], or the scaled speech intelligiblity index (SII) [22], [23].

While the sigmoid $A_k(\xi_k)$ differs from the heuristically selected curves used in AI and SII, the similarity is well within the precision of the reasoning used to arrive at the AI and SII formulation. Thus, (8) forms a theoretical justification for this classical work on speech intelligibility.

III. Optimizing Information Throughput

Our objective is to optimize the effectiveness of the communication process by selecting a good enhancement operator $\mathcal{G}$. Let us consider a common time-frequency representation such as that obtained with a paraunitary Gabor or DCT filterbank. For this representation, the assumption of a memoryless stationary process is reasonable. We consider a memoryless linear and time-invariant operator $(\mathcal{G}(X))_k = \sqrt{b_k}X_k$, which is affine, and redistributes signal power by multiplying each frequency channel with a gain $\sqrt{b_k}$. The redistribution is subject to an overall signal power preservation constraint.

The intelligibility optimization problem is now

$$\begin{aligned}&\max_{\{b_k\}} I(S;\tilde{Z})\\ &\text{subject to} \sum_{k\in\kappa} b_k\sigma^2_{X_k} - B = 0 \text{ and } b_k \geq 0, \forall_k,\end{aligned} \tag{9}$$

where B is the power of the vector X. The problem can be solved using the Karush-Kuhn-Tucker (KKT) conditions.

While the correlation coefficients $\rho_{S_kX_k}$ and $\rho_{Y_kZ_k}$ are fixed, the correlation coefficient $\rho_{\tilde{X}_k\tilde{Y}_k}$ varies with the coefficient b_k as follows:

$$\rho_{\tilde{X}_k\tilde{Y}_k} = \frac{1}{\sqrt{1+\frac{\sigma^2_{N_k}}{b_k\sigma^2_{X_k}}}}. \tag{10}$$

Denoting $\rho_{\tilde{X}_k\tilde{Y}_k} = \frac{1}{\sqrt{1+\frac{\sigma^2_{N_k}}{b_k\sigma^2_{X_k}}}}$., the objective is

$$\begin{aligned}&\max_{\{b_k\}} \sum_{k\in\kappa}\frac{1}{2}\log\left(\frac{b_k\sigma^2_{X_k}+\sigma^2_{N_k}}{(1-\rho^2_{0,k})b_k\sigma^2_{X_k}+\sigma^2_{N_k}}\right)\\ &\text{subject to} \sum_{k\in\kappa} b_k\sigma^2_{X_k} - B = 0 \text{ and } b_k \geq 0, \forall_k,\end{aligned} \tag{11}$$

which is a convex optimization problem as the objective function is concave. From (11) we construct the Lagrangian

$$\begin{aligned}&\mathcal{L}(\{b_k\},\lambda,\{\mu_k\}) =\\ &\sum_{k\in\kappa}\frac{1}{2}\log\left(\frac{b_k\sigma^2_{X_k}+\sigma^2_{N_k}}{(1-\rho^2_{0,k})b_k\sigma^2_{X_k}+\sigma^2_{N_k}}\right) + \lambda b_k\sigma^2_{X_k} + \mu_k b_k.\end{aligned} \tag{12}$$

The μ_k are nonnegative and λ is nonpositive (as the mutual information is monotonically increasing as a function of b_k).

Differentiating the Lagrangian to the b_k and setting the results to zero leads to the stationarity conditions of the KKT conditions:

$$\begin{aligned}0 = &\frac{1}{2}\frac{\sigma^2_{X_k}}{b_k\sigma^2_{X_k}+\sigma^2_{N_k}}\\ &-\frac{1}{2}\frac{(1-\rho^2_0)\sigma^2_{X_k}}{(1-\rho^2_0)b_k\sigma^2_{X_k}+\sigma^2_{N_k}} + \lambda\sigma^2_{X_k} + \mu_k, \forall_k.\end{aligned} \tag{13}$$

Multiplying by the denominators leads to a quadratic in b_k:

$$\alpha b_k^2 + \beta b_k + \gamma = 0 \tag{14}$$

with

$$\gamma = \frac{1}{2}\rho^2_{0,k}\sigma^2_{X_k}\sigma^2_{N_k} + (\lambda\sigma^2_{X_k}+\mu_k)\sigma^4_{N_k}, \tag{15}$$

$$\beta = (\lambda\sigma^2_{X_k}+\mu_k)(2-\rho^2_{0,k})\sigma^2_{X_k}\sigma^2_{N_k}, \tag{16}$$

$$\alpha = (\lambda\sigma^2_{X_k}+\mu_k)(1-\rho^2_{0,k})\sigma^4_{X_k}. \tag{17}$$

Let us study the behavior of the quadratic (14). It is guaranteed to have real roots if $\beta^2 - 4\alpha\gamma \geq 0$. e consider what happens when $\mu_k = 0$. First we notice that $4\alpha\gamma$ consists of two terms: $\frac{1}{2}\rho^2_0\sigma^2_{X_k}\sigma^2_{N_k}\alpha$, which is negative for and $(\lambda\sigma^2_{X_k}+\mu_k)\sigma^4_{N_k}\alpha$, which is positive for $\mu_k = 0$. If the latter term is smaller than β^2 we have that b_k has real roots; this is true if

$$4(1-\rho^2_{0,k}) \leq (2-\rho^2_{0,k})^2, \tag{18}$$

which is always true as $\rho^2_{0,k} \in [0,1]$. The roots may, however, both be negative and in this case the term $\mu_k b_k$ must be sufficiently negative to force the root to $b_k = 0$. This leads to the standard KKT solution. A simple line search algorithm for the λ that provides the correct overall power is:

(1) select λ;
(2) solve (14) with $\mu_k = 0$ for all b_k;
(3) set any negative b_k to zero;
(4) check if the power $\sum_{k\in\kappa} b_k\sigma^2_{X_k}$ is sufficiently close to the desired overall power B. If not, then adjust λ to be more negative if the power is too high and more positive if the power is too low.

The algorithm is easily extended to a bi-section algorithm.

It can now be seen that, in contrast to the case where the production and interpretation noise are not considered, increasing a single $\sigma^2_{N_k}$ can either decrease or increase b_k. From the standard quadratic root formula it follows that for a given ρ^2_0 and $\sigma^2_{X_k}$ the change in value for b_k depends on the term $-4\gamma\alpha$ in the root. Consider again $\mu_k = 0$. The behavior depends on whether the positive term $-\frac{1}{2}\rho^2_0\sigma^2_{X_k}\sigma^2_{N_k}\alpha$ or the negative term $-(\lambda\sigma^2_{X_k}+\mu_k)\sigma^4_{N_k}\alpha$ is larger. The first term being larger corresponds to the “saturated” case discussed at the end of the introduction and the case where the second terms is larger to the “unsaturated” case.

IV. Results

In this section we provide both illustrative results that provide insight in how the algorithm works, and the results of a formal listening test. We contrast mutual information for models with and without observation and interpretation noise and also compare our results to the state-of-the-art.

The experiments were performed on 16 kHz sampled speech and frequency dependent gains were implemented with a Gabor analysis and synthesis filterbanks with oversampling by a factor two and a Fourier transform size of 512 and a square-root Hann window. Note that while the selected gains may result in the processed complex signal not to be in the space spanned by the forward transform, the inverse Gabor implicitly performs an orthonormal (i.e., optimal) projection onto that space. To obtain

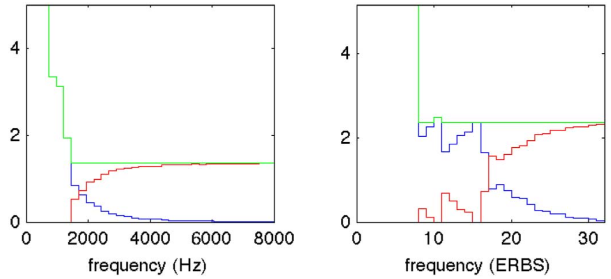

Fig. 1. Optimization of mutual information: power of enhanced signal $\sigma^2_{\tilde{X}_k}$ (red), noise signal $\sigma^2_{N_k}$ (blue), and their sum (green). Linear scale (left) and ERB scale (right) are shown.

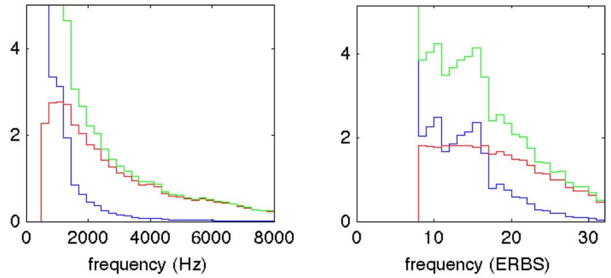

Fig. 2. Optimization of mutual information with production and interpretation noise: power of enhanced signal $\sigma^2_{\tilde{X}_k}$ (red), noise signal $\sigma^2_{N_k}$ (blue), and their sum (green). Linear scale (left) and ERB scale (right) are shown.

the auditory representation, 64 gammatone filters were used, uniformly distributed on the ERB scale.

The illustrative figures show the results for an eight-second utterance spoken by a German male speaker with a noise that was recorded in a train. The channel SNR for the examples in the figures is -5 dB, measured over the entire utterance and the value $\rho_{0,k} = 0.2$ for all bands.

For the listening experiments we used speech-shaped noise. In this case the values for $\rho_{0,k}$ were computed from the band-importance tables in the SII standard [22]. Only the auditory domain optimization version of the algorithm was used in the listening experiments. Nine native Dutch speakers listened to 96 five-word sentences created from a closed set of words and had to select each word from a set of 10 [24].

Fig. 1 shows results for the maximization of the mutual information between S and Z for the case of zero production and observation noise ($\rho_{0,k} = 1$). The left figure is for optimization in the linear frequency domain and the right figure for the auditory representation case. The results correspond to the standard waterfilling solution of communication theory (e.g., [17]). It is seen that for the higher frequency bands, the optimal gains b_k for each band k of the observable signal X_i are selected to make $\sigma^2_{\tilde{X}_k} + \sigma^2_{N_k}$ constant.

Importantly, it can be observed that for this type of noise (and commonly for most noise types), the channel SNR in the high frequency bands is high. If $\rho_{0,k} < 1$ and the production SNR is lower than the channel SNR in these frequency bands, and if a power constraint applies, then resources are not used effectively. In other words, the signal intelligibility would not be reduced if the signal power would be reduced in these bands. Thus, this power can be spent elsewhere.

Fig. 2 shows what happens to the scenarios of Fig. 1 if the production and interpretation SNR are considered (the figures are on the same scale). As mentioned, we set $\rho_{0,k} = 0.2$ for all k. It is seen that for the higher frequency bands, the power $\sigma^2_{\tilde{X}_k} = b_k \sigma^2_{X_k}$ is essentially proportional the noise power $\sigma^2_{N_k}$. This allows more of the signal energy to be used in the lower energy bands as compared to Fig. 1.

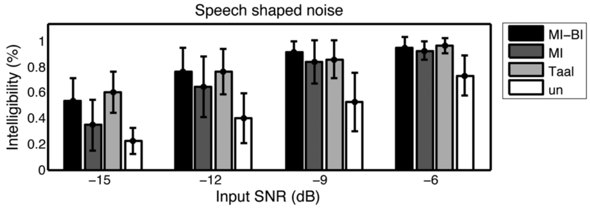

Fig. 3. Listening test results.

The listening test results shown in Fig. 3 confirm that the illustrative results of Fig. 2 correspond to an improvement in intelligibility. The figure shows results for unprocessed speech (Un), mutual-information optimization (MI), and mutual information optimization considering production and interpretation noise (MI-B). Additionally it shows the results for the reference state-of-the-art result of Taal *et al.* [10]. For a significance level of $\alpha = 0.05$, all processed speech is more intelligible than unprocessed speech, except MI at -12 dB. For -12 dB and -15 dB, MI-B is significantly more intelligible than MI. Thus, consideration of production and interpretation noise improves intelligibility when using mutual information as criterion. The differences between MI-B and the reference are not statistically significant. This is to be expected as *i*) the reference is based on the SII relation (7) (in contrast to MI-B, the reference uses a heuristically derived weighting function) *ii*) in this first experiment we used $\rho_{0,k}$ that were computed from the band importance function I_k of the SII standard, which is also used by the reference.

V. Conclusion

A simple information-theory based model of speech communication suffices for state-of-the-art enhancement of the intelligibility of speech played out in a noise environment. The model makes the plausible assumption that both the production and the interpretation process in the speech communication chain are subject to noise that scales with the signal level.

The model suggests that the impact of the noise in the production and interpretation processes is similar. If production and interpretation fidelity have increasing marginal cost, then similar signal-to-noise ratios for the production and interpretation processes would minimize overall cost. Moreover, our model suggests that it is reasonable to surmise that the average spectral density of speech matches typical noise in the environment.

Our approach can be refined in a number of aspects. Regularization can be applied to reduce intelligibility enhancement if no noise is present. Other distributions than the Gaussian distribution can be used for the speech. In the subjective experiments, we used fixed or SII-standard derived settings for the production and interpretation noise. Instead, one can use direct measurements of the variability of the observable speech signal for a given set of utterances. The simple enhancement operator can be replaced by more effective nonlinear enhancement methods.

References

[1] P. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[2] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook-based Bayesian speech enhancement for nonstationary environments,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, pp. 441–452, Feb. 2007.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain based single-microphone noise reduction for speech enhancement-a Survey of the State of the Art*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[4] V. Grancharov, J. H. Plasberg, J. Samuelsson, and W. B. Kleijn, “Generalized postfilter for speech quality enhancement,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 1, pp. 57–64, Jan. 2008.

[5] R. Niederjohn and J. Grotelueschen, “The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression,” *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 277–282, 1976.

[6] B. Sauert, G. Enzner, and P. Vary, “Near end listening enhancement with strict loudspeaker output power constraining,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, 2006.

[7] B. Sauert and P. Vary, “Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 1919–1923.

[8] T. C. Zorilă, V. Kandia, and Y. Stylianou, “Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression,” in *ISCA Interspeech*, Portland, OR, USA, 2012.

[9] P. N. Petkov, G. E. Henter, and W. B. Kleijn, “Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 5, pp. 1035–1045, 2013.

[10] C. Taal, J. Jensen, and A. Leijon, “On optimal linear filtering of speech for near-end listening enhancement,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, 2013.

[11] J. B. Crespo and R. C. Hendriks, “Multizone speech reinforcement,” *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 54–66, 2014.

[12] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA, USA: MIT Press, 1997.

[13] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[14] A. Leijon, “Articulation index and Shannon mutual information,” in *Hearing From Sensory Processing to Perception*, B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, Eds. Berlin, Germany: Springer-Verlag, 2007.

[15] J. Taghia and R. Martin, “Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, pp. 6–16, Jan. 2014.

[16] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. i. Model structure,” *J. Acoust. Soc. Amer.*, vol. 99, pp. 3615–3622, 1996.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.

[18] B. Moore and B. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” *J. Acoust. Soc. Amer.*, vol. 74, pp. 750–753, 1983.

[19] J. O. Smith and J. S. Abel, “Bark and ERB bilinear transform,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, Nov. 1999.

[20] N. R. French and J. C. Steinberg, “Factors governing the intelligibility of speech sounds,” *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[21] K. D. Kryter, “Methods for the calculation and use of the articulation index,” *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[22] “American national standard methods for calculation of the speech intelligibility index,” ANSI/ASA S3.5-1997 (R2012) 2012.

[23] K. S. 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,” *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[24] R. Houben, J. Koopman, H. Luts, K. Wagener, A. van Wieringen, H. Verschuure, and W. Dreschler, “Development of a Dutch matrix sentence test to assess speech intelligibility in noise,” *Int. J. Audiol.*, pp. 1–4, 2014, epub ahead of print.