International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Simple Model of Speech Communication and its Application to Intelligibility Enhancement

Dr.K.B.S.D.Sarma & Pravin Akula
1,2Professor, Dept of E.C.E, BVC Engineering College, Odalarevu

***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 environ-ment. State- of-the-art experimental results confirm the practical value of the model.*

I. Introduction:

Modern communication technology allows a user to communicate from almost anywhere to almost any-where. 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 anal-ysis 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 rep-resented by a K-dimensional stationary discrete-time random process. The process is composed of real or complex scalar vari-ables $S_{k,i}$, 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

(1) $$X_{k,i} = S_{k,i} + V_{k,i},$$

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

(2) $$Y_{k,i} = X_{k,i} + N_{k,i}$$

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

(3) $$Z_{k,i} = Y_{k,i} + W_{k,i},$$

where $W_{k,i}$ is “interpretation” noise. Note that $S \rightarrow X \rightarrow Y \rightarrow Z$ 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 memory less, which is reasonable for time frequency signal representations.

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

B. Tractable Model that Includes Enhancement

We now insert a machine-based enhancement operator in the Markov chain. If we mark by all signals affected by the enhancement operator we get a Markov chain S→X→X→Ỹ→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 from here-on forward. For the Gaussian case it can be shown that

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

Next, we consider how the theory is affected if the signal is
interpreted in its 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.

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

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

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

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

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

If we identify I_k as the scaled *band- importance function* and $A_k(.)$ as the *weighting function* the mutual information can be interpreted as the scaled articulation index (AI) or the scaled speech intelligiblity index (SII) .While the sigmoid $A_k(\xi_k)$
differs from the heuristically se-lected 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

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

(8)

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 and are fixed,
the correlation coefficient varies with the coefficient
as follows:

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

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

(9)

which is a convex optimization problem as the objective function
is concave. From (9) we construct the Lagrangian

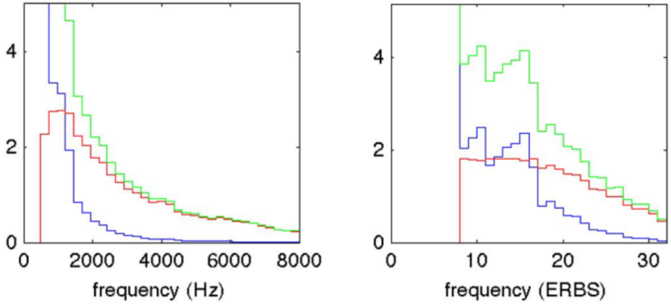
$$\mathcal{L}(\{b_k\},\lambda,\{\mu_k\}) = \sum_{k\in\kappa}\frac{1}{2}\log\left(\frac{b_k\sigma^2_{X_k}+\sigma^2_{N_k}}{(1-\rho^2_{0,k})b_k\sigma^2_{X_k}+\sigma^2_{N_k}}\right)+\lambda b_k\sigma^2_{X_k}+\mu_k b_k. \qquad \text{-(10)}$$

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 . From the standard quadratic root formula it follows that for a given ρ_0^2 and $\sigma^2_{N_k}$ the change in value for depends on the term in
the root.

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

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



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

Fig. 1 shows results for the maximization of the mutual information between and for the case of zero production and observation noise ( ). 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 .It is seen that for the higher frequency bands, the optimal gains for each band of the observable signal are selected to make $\sigma^2_{\hat{X}_k}+\sigma^2_{N_k}$ constant.

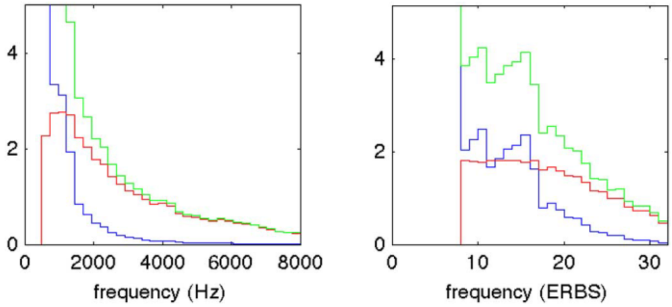


Fig. 2. Optimization of mutual information with production and interpretation noise: power of enhanced signal $\sigma^2_{\hat{X}_k}$ (red), noise signal $\sigma^2_{N_k}$ (blue), and their sum (green). Linear scale (left) and ERB scale (right) are shown.

Fig. 2 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 for all . It is seen that for the higher frequency bands, the power is essentially proportional the noise power . This allows more of the signal energy to be used in the lower energy bands as compared to Fig. 1.

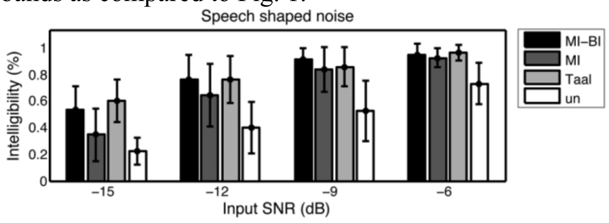


Fig. 3. Listening test results.

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). 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 (in contrast to MI-B, the reference uses a heuristically derived weighting function) *ii*) in this first experiment we used that were computed from the band importance function 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.

AUTHOR DETAILS

Dr.K.B.S.D.SARMA working as a Professor in the Department of Electronics and Communication Engineering, BVC Engineering College. Odalarevu. He is published 15 paper publications in various national and International journals and conferences. His research interest includes Signal Processing and VLSI System designs and its Applications.

Dr.K.B.S.D.SARMA
Professor,
BVC Engineering College, Odalarevu
bsdsarma.kompella007@gmail.com

Pravin Akula obtained his B.E Degree in Electronics and Communications Engineering from SRKR Engineering college and M.Tech degree in Instrumentation and Control systems from JNTUK, Kakinada. He is presently working as Professor in the Department of Electronics and Communication Engineering, B V C Engineering College, Odalarevu. He is presently pursuing Ph.D degree in K L University. Vijayawada, A.P India. He has 15 paper publications in various national and International journals and conferences. His current research interests includes VLSI Low Power . He is the Life Member of MISTE.