Optimizing Speech Intelligibility in a Noisy Environment

W. Bastiaan Kleijn *Fellow, IEEE,* João B. Crespo *Student Member, IEEE,*
Richard C. Hendriks *Member, IEEE,* Petko N. Petkov *Student Member, IEEE,*
Bastian Sauert, Peter Vary *Fellow, IEEE*

Abstract

Modern communication technology facilitates communication from-anywhere to-anywhere. As a result, low speech intelligibility has become a common problem, which is exacerbated by the lack of feedback to the talker about the rendering environment. In recent years, a range of algorithms has been developed to enhance the intelligibility of speech rendered in a noisy environment. We describe methods for intelligibility enhancement from a unified vantage point. Before one defines a measure of intelligibility, the level of abstraction of the representation must be selected. For example, intelligibility can be measured on the message, on the sequence of words spoken, on the sequence of sounds, or on a sequence of states of the auditory system. Natural measures of intelligibility defined at the message level are mutual information and the hit-or-miss criterion. The direct evaluation of high-level measures requires quantitative knowledge of human cognitive processing. Lower level measures can be derived from higher level measures by making restrictive assumptions. We discuss the implementation and performance of some specific enhancement systems in detail, including speech intelligibility index (SII) based systems and systems aimed at enhancing the sound-field where it is perceived by the listener. We conclude with a discussion of the current state of the field and open problems.

I. Introduction

Humans adapt their speech to the physical environment. Based on the facial expression of the listener, a talker may repeat or reformulate the message. A noisy environment gives rise to the Lombard effect, e.g., [1], an involuntary change in the speech characteristics that makes speech more intelligible.

In modern communication systems, the talker often has little or no awareness of the physical environment that the speech is rendered in. This is perhaps most obvious for current-generation speech synthesis,

W. B. Kleijn is with Victoria University of Wellington, New Zealand and Delft University of Technology, the Netherlands, P. Vary is with RWTH Aachen University, Germany, B. Sauert was with RWTH Aachen University, Germany and is now with HEAD acoustics, Germany, R.C. Hendriks and J.B. Crespo are with Delft University of Technology, the Netherlands, P. N. Petkov is with KTH, the Royal Institute of Technology, Stockholm, Sweden.

which produces speech without consideration of the rendering environment. It is also a major factor in human-to-human communications as communication technology degrades or severes the auditory and visual links between the talker and the environment. For example, an announcer at a railway station generally receives little visual or auditory feedback. Similarly, a phone user lacks information about the rendering environment, even less so if effective noise-suppression technology is used.

The lack of feedback, together with the recent ability to communicate from-anywhere to-anywhere often leads to low intelligibility. Phonebooths are a relic of the past: the mobile phone is expected to function in any environment, whether it is a car, a cafeteria, or a windstorm. Thus, there is a strong motivation for algorithms that can improve the intelligibility of speech rendered in a noisy environment.

Ever since the early work of Griffiths [2] and Niederjohn and Grotelueschen [3], researchers have attempted to create processing methods that increase the intelligibility of speech in a noisy environment. Driven by the rapid growth of mobile telephony, research efforts on intelligibility in noise have increased significantly in the last five years. The result is that it is now possible to increase the intelligibility of speech in noise significantly, e.g., [4]–[11]. Approaches to intelligibility enhancement are increasingly based on the mathematical optimization of quantitative measures that are hypothesized to represent intelligibility accurately. First introduced by [2], the optimization approach has been used in numerous recent studies, starting with [12]. The optimization criteria vary widely as the signal processing algorithms are derived from different viewpoints and with different computational and delay constraints. Criteria used include the probability of correct phoneme recognition [11], auditory models [6], [13], [14], the articulation index [2], the speech intelligibility index [4], [8], mutual information [15], and sound-field distortion [16].

In this tutorial, we describe a range of methods for intelligibility enhancement from a unified vantage point, delineating the similarities and dissimilarities between the various approaches. In contrast to the broad overview of human and algorithmic modifications that affect intelligibility in [7], our discussion focuses on the definition and use of quantitative measures of intelligibility, showing that many of these measures can be derived from the same basic principle. A number of quantitative measures is defined in section II. We then discuss three different implementations in some detail in section III and end with conclusions and a view of the future in section IV.

II. Measures of Intelligibility

A. Defining Intelligibility

The word *intelligibility* expresses a qualitative measure of whether a conveyed message is interpreted correctly by a human listener. To define quantitative instrumental measures of intelligibility we must select a level of abstraction. That is, we must decide if we measure intelligibility on the sequence of words spoken, on the sequence of sounds, on a sequence of states of the auditory system, or on the acoustic signal waveform. A word sequence is an example of a description at a high level of abstraction whereas a signal waveform is a description at a low level of abstraction.

The higher the level of abstraction, the more fundamental the measure of intelligibility: the objective of speech is to convey a message and not to convey a sequence of sounds. A particular measure will be useful for enhancement at its own level of abstraction and below. Consider an intelligibility measure operating at the word sequence level. It can be used to evaluate which of a set of sentence formulations with similar meaning is more intelligible. It can also be used to evaluate if a particular spectral modification (e.g., a particular filtering operation) makes speech more intelligible.

The generality of high-level measures has a cost: we must map the observations into a sequence at that high abstraction level. For acoustic observations and a measure operating at the word-sequence level this requires a robust model of hearing that maps the observed acoustic signal into a word sequence. Therefore, although it cannot optimize linguistic formulations, an intelligibility measure operating on a sequence of auditory states may be attractive when optimizing a spectral modification of the signal.

While illusive in practical measurements, the message itself, a random variable that we denote as $\mathbf{M}$, can be used to define the most basic measure of intelligibility. (To aid clarity, we will write random variables as bold-face characters and their realizations as regular characters.) In the following we will show how such a basic measure can be used to derive measures that have been derived earlier on a heuristic basis. To facilitate our reasoning, we will be opportunistic and sometimes describe the messages as countable, which is consistent with the notion that a message is a discrete word sequence, and at other times as continuous, which is consistent with the notion that articulation is continuously variable. To avoid confusion, we add a breve, as in $\breve{\mathbf{M}}$, whenever messages are considered countable.

A natural measure of intelligibility is the mutual information between the message conveyed by the talker $\breve{\mathbf{M}}_T$ and the message interpreted by the listener $\breve{\mathbf{M}}_L$:

$$I(\breve{\mathbf{M}}_L;\breve{\mathbf{M}}_T) = \sum_{\breve{M}_L,\breve{M}_T} p_{LT}(\breve{M}_L,\breve{M}_T) \log \frac{p_{L|T}(\breve{M}_L|\breve{M}_T)}{p_L(\breve{M}_L)}, \tag{1}$$

where we used the simplified notation $p_{LT} = p_{\breve{\mathbf{M}}_L \breve{\mathbf{M}}_T}$ and $p_{L|T} = p_{\breve{\mathbf{M}}_L|\breve{\mathbf{M}}_T}$ for the joint and conditional probabilities and use the same convention for the marginal probabilities of the conveyed and received messages and p_T and p_L.

We can reformulate the criterion (1) as a measure of distortion $D(\breve{\mathbf{M}}_L, \breve{\mathbf{M}}_T)$ that is a functional of $p_{L|T}$. Mutual information is nonnegative and cannot be larger than the entropy $H(\breve{\mathbf{M}}_T)$. Thus, the difference $D(\breve{\mathbf{M}}_L, \breve{\mathbf{M}}_T) = H(\breve{\mathbf{M}}_T) - I(\breve{\mathbf{M}}_L; \breve{\mathbf{M}}_T)$ is nonnegative and can be interpreted as a distortion. It can be written as a general distortion measure operating on $p_{L|T}$ for a given talker message distribution p_T:

$$D(\breve{\mathbf{M}}_L, \breve{\mathbf{M}}_T) = \sum_{\breve{M}_T} p_T(\breve{M}_T) \sum_{\breve{M}_L} d(p_{L|T}(\breve{M}_L|\breve{M}_T)), \tag{2}$$

where d is a nonnegative function of $p_{L|T}(\breve{M}_L|\breve{M}_T)$. For the mutual information based distortion measure $d(p_{L|T}(\breve{M}_L|\breve{M}_T)) = p_{L|T}(\breve{M}_L|\breve{M}_T)\log(p_L(\breve{M}_L)/p_{L,T}(\breve{M}_L, \breve{M}_T))$, where we note that the argument of the logarithm can be written in terms of $p_{L|T}(\breve{M}_L|\breve{M}_T)$ and the given $p_T(\breve{M}_T)$ only. The intelligibility enhancement problem is to find the $p_{L|T}$ that minimizes the distortion (2) subject to the constraints set by the scenario.

An alternative to the mutual information based distortion measure can be based on the hit-or-miss distortion, $d(p_{L|T}(\breve{M}_L|\breve{M}_T)) = p_{L|T}(\breve{M}_L|\breve{M}_T)(1-\delta_{\breve{M}_L,\breve{M}_T})$, where $\delta_{\breve{M}_L,\breve{M}_T}$ is a Kronecker delta function. In this case (2) becomes

$$D(\breve{\mathbf{M}}_L, \breve{\mathbf{M}}_T) = 1 - \sum_{\breve{M}_T} p_{LT}(\breve{M}_T, \breve{M}_T) = 1 - \mathrm{E}_T[p_{L|T}(\breve{\mathbf{M}}_T|\breve{\mathbf{M}}_T)]. \tag{3}$$

The conditional probability $p_{L|T}(\breve{M}_T|\breve{M}_T)$ in (3) corresponds to *the probability that the message is interpreted correctly*. Thus, an alternative to maximizing the mutual information of the conveyed and received message is to maximize the expected probability of correct message interpretation, $\mathrm{E}_T[p_{L|T}(\breve{\mathbf{M}}_T|\breve{\mathbf{M}}_T)]$, where the expectation is over the conveyed messages, $\breve{\mathbf{M}}_T$. We will discuss the practical use of this high-level measure in section II-B1.

While the measures (1) and (3) are general, they cannot be used directly. Either the description of the message, or the human cognitive system must be approximated such that the measures can be applied to observable signals. The paradigm shows where such approximations are made, but it does not show their quantitative impact. Thus, experiments must be used to verify the validity of the resulting system.

Next, we consider how to derive a low-level, acoustics-based measure from a high-level, message-based measure. For this it is convenient to consider the message as a continuous variable. A conveyed speech message $\mathbf{M}_T$ is rendered in the form of an acoustic signal, which we represent by an acoustic sequence $\mathbf{a}_T$. The sequence $\mathbf{a}_T$ can, for example, consist of signal samples or short-term spectral descriptions, such as cepstral vectors. This sequence is rendered in a noisy environment and the listener observes a corrupted sequence $\mathbf{a}_L$, which is then interpreted as a message $\mathbf{M}_L$. The communication process thus forms a Markov chain $\mathbf{M}_T \rightarrow \mathbf{a}_T \rightarrow \mathbf{a}_L \rightarrow \mathbf{M}_L$. It is natural that environmental noise makes the mapping $\mathbf{a}_T \rightarrow \mathbf{a}_L$ stochastic.

Upon reflection, it is clear that the mappings $\mathbf{M}_T \rightarrow \mathbf{a}_T$ and $\mathbf{a}_L \rightarrow \mathbf{M}_L$ are also stochastic: a message is generally not formulated and never articulated in precisely the same manner, and the interpretation of the acoustic sequence $\mathbf{a}_L$ is subject to random variations during the human cognitive process. Anticipating the discussions in Section II-B1, it can be argued that these variations are captured by the statistical modeling of modern automatic speech recognition (ASR) algorithms. If we assume the message formulation is perfect, a simple but effective model of the production and interpretation processes is that they are subject to additive noise components [15], which we will refer to as, respectively, *production noise* and *interpretation noise*. For example, variability in articulation across different persons may be approximated as additive noise in a representation based on cepstral or log spectral vectors.

For convenience let us define auxiliary bijective mappings $M_T \leftrightarrow s_T$ and $M_L \leftrightarrow s_L$, where s_T and s_L are realizations of random acoustic sequences. We have

$$\begin{aligned} \mathbf{a}_T &= \mathbf{s}_T + \mathbf{v}_T \\ \mathbf{a}_L &= \mathbf{a}_T + \mathbf{v}_E \\ \mathbf{s}_L &= \mathbf{a}_L + \mathbf{v}_L \end{aligned} \tag{4}$$

where $\mathbf{v}_T$, $\mathbf{v}_E$, and $\mathbf{v}_L$ are additive noise processes, modeling the production noise, environmental noise and interpretation noise, respectively. Note that the system model differs from the standard system model in communication theory, which does not include production noise and interpretation noise.

To facilitate analysis, let us assume the sequences $\mathbf{s}_T$, $\mathbf{v}_T$, $\mathbf{v}_E$ and $\mathbf{v}_L$ to be jointly Gaussian processes. Furthermore, we denote by $\rho_{\mathbf{sa}}$ the correlation coefficient of (the samples of the) processes $\mathbf{s}$ and $\mathbf{a}$ and write $\rho_0 = \rho_{\mathbf{s}_T\mathbf{a}_T}\rho_{\mathbf{a}_L\mathbf{s}_L}$. Let us first consider the case where the signals are white. Exploiting that mutual information is invariant under reparametrization of the marginal variables, it is then easy to see that [15]

$$I(\mathbf{M}_L; \mathbf{M}_T) = I(\mathbf{s}_T; \mathbf{s}_L) = -\frac{1}{2} \log \frac{(1-\rho_0^2)\xi + 1}{\xi + 1}, \tag{5}$$

where $\xi = \frac{\sigma^2_{\mathbf{a}_T}}{\sigma^2_{\mathbf{v}_E}}$ is the SNR of the acoustic channel $\mathbf{a}_T \rightarrow \mathbf{a}_L$, and $\sigma^2_{\mathbf{a}_T}$ and $\sigma^2_{\mathbf{v}_E}$ are the variances of processes $\mathbf{a}_T$ and $\mathbf{v}_E$, respectively. An important and intuitive conclusion that can be drawn from (5) is that if the environmental noise variance is small compared to the production and interpretation noise variances, then the mutual information between talker and listener is not affected significantly by the environmental noise.

The spectral coloring of the acoustic content can be accounted for by splitting the signal into spectral bands such that each band can be approximated as white. If we assume the signals to be stationary, the frequency bands are independent and the mutual information can be written as the sum of the mutual informations in the bands:

$$I(\mathbf{M}_L;\mathbf{M}_T) = -\frac{1}{2}\sum_i \log \frac{(1-\rho_{0,i}^2)\xi_i+1}{\xi_i+1}, \tag{6}$$

where i is the band index and where $\xi_i = \frac{\sigma^2_{\mathbf{a}_{T,i}}}{\sigma^2_{\mathbf{v}_{E,i}}}$ is the SNR of the acoustic channel in band i. Note that the SNR in (6) is computed on whichever representation is used for the acoustic features. Also note that the variances $\sigma^2_{\mathbf{a}_{T,i}}$ and $\sigma^2_{\mathbf{v}_{E,i}}$ are generally unknown and must be estimated in practice. For example, if the acoustic features are based on short-time DFT coefficients, variance estimation can be based on the short-time DFT periodogram, i.e., $\hat{\sigma}^2_{\mathbf{a}_{T,i}} = |a_{T,i}|^2$ having a variance of $E\left[|\mathbf{a}_{T,i}|^2\right]^2$. The low-level measure (6) can then be used directly to optimize speech intelligibility [15].

The frequency resolution of the human auditory system decreases with frequency, which reduces the mutual information from that obtained with (6) for a uniform high resolution. An improved model of information transfer is obtained by assuming that the signal is represented with one independent component per equivalent rectangular bandwidth (ERB), which is consistent with studies on intelligibility [17]. We show in section II-B3 that this approach provides an information-theoretical justification of the well-known speech intelligibility index (SII, see also Fig. 2) [18], a low-level measure of intelligibility.

B. Practical Measures of Intelligibility

Existing practical measures of intelligibility generally operate at the word-sequence level, at the level of a sequence of auditory states, or at the level of short-term spectra. We discuss these classes below. We end this subsection with a discussion of the constraints that must be imposed on the optimization.

1) Measures Operating on a Word-Sequence: In section II-A, we discussed that the expected probability of correct interpretation of the message, $\mathrm{E}_T[p_{L|T}(\breve{\mathbf{M}}_T|\breve{\mathbf{M}}_T)]$, is a reasonable measure of intelligibility. This measure can be approximated as $\overline{p_{L|T}(\breve{M}_T|\breve{M}_T)}$ on real-world data, where the overbar indicates averaging over realizations $\breve{M}_T$. If the averaging is done in time, i.e., over segments of a single larger message (e.g., words), then this operation assumes ergodicity. The measure is easily evaluated in a test with human test subjects, where $p_{L|T}(\breve{M}_T|\breve{M}_T)$ can be estimated using histograms. A machine-based quantitative measure requires a mapping from any particular acoustic observation a_L to a message $\breve{M}_L$

that captures the probabilistic nature of this mapping as performed by humans. As will be discussed in Section III-B2, the standard approach to ASR computes the probability of the observations given a message (word, or word sequence). The basic assumption for machine-based intelligibility enhancement is then that the trend of ASR word-probability in noise tracks the trend of human recognition performance in noise sufficiently well for the modification parameters that are optimized. Experiments confirmed this hypothesis [11], [19] for a particular set of practical systems.

2) Measures Operating on a Sequence of Auditory States: Particularly if the types of modification are restricted, it is advantageous to minimize the delay and computational requirements of the intelligibility measure. Let us assume that the modification is a spectral modification, that the word sequence and speaking rate are fixed, and that the highest intelligibility is achieved by the original speech without environmental noise. (The latter assumption is an additional simplification required for this approach.) Then it is natural to use a distortion measure operating on the sequence of auditory states as a measure of intelligibility. Such measures can exploit that quantitative knowledge of the auditory periphery has increased significantly in the last three decades (e.g., [20]).

The straight comparison of the auditory states of the conveyed and received signal ignores the production noise $\mathbf{v}_T$ of (4). That is, the auditory model does not weigh signal components according to their relevance in terms of precision of signal production. However, the auditory model precision of a speech component may form a reasonable match to the precision of speech production, simplifying the introduction of production noise.

Although auditory models differ in how exactly the inner ear representation is obtained, they follow in many cases a similar strategy for modeling the auditory system. In Fig. 1 we outline the basic building blocks of the psycho-acoustic model presented in [21], which is simple but representative of many other models, such as [20]. The first stage of the auditory model consists of a filter that mimics the frequency characteristics of the outer and middle ear. This filter is cascaded with an auditory filterbank that models processing at the level of the basilar membrane in the cochlea. Subsequently, the envelope of each of the outputs of the auditory filters is obtained, which simulates the transduction of the inner hair-cells. To model an absolute hearing threshold, a constant is added to each envelope. In the current context, this threshold corresponds to an interpretation noise. In the final stage a log transform is used to model the loudness dependent compression of the auditory filterbank outputs by the outer hair-cells. An important difference between the model from [21] and the more advanced model presented in [20] is the logarithmic transform, which is a simplification of the adaptation loops that are used in [20]. The simplification particularly affects the output near transitions, where the gain of adaptation loops changes.

By applying an auditory model to the acoustic sequences a_T and a_L and comparing the results, a distortion measure can be obtained. Mutual information is a natural measure for this purpose, but, to our best knowledge, it has not been applied to the auditory representation for intelligibility enhancement. Note that while mutual information is not affected by smooth invertible mappings, auditory representations likely are not smooth mappings from features such as cepstra, or line spectral frequencies. This suggests that it may be essential to consider the detailed behavior of more sophisticated auditory models.

In the literature, various measures have been used to compare the auditory representations of a_T and a_L. In [14], it was shown that an ℓ_1 criterion leads to a mathematically tractable method and to provide good results for intelligibility enhancement. [13] uses a similar auditory model for the so-called *glimpse proportion* measure of intelligibility: rather than comparing a_T and a_L directly, it compares the auditory representation of the a_T with the auditory representation of the environmental noise v_E. The glimpse proportion approach computes the proportion of signal blocks where the auditory representation of the signal is louder than the noise. In more recent work on the glimpse proportion, a sigmoidal function is applied to the difference of the auditory signal and noise representations [6], [22]. The method provides good intelligibility enhancement [6], [22], [23]. Both the ℓ_1 criterion and glimpse proportion approaches do not explicitly consider the information conveyed in a particular signal component, which should, at least in principle, be a disadvantage compared to mutual information based approaches.

3) Measures Operating on Spectral Band Powers: The mutual information between $\mathbf{M}_L$ and $\mathbf{M}_T$ (6) can be seen to correspond to a classic view of intelligibility based on band powers of the auditory filter bank [17], [18], [24]–[27] by writing it as

$$I(\mathbf{M}_L;\mathbf{M}_T) = \sum_i \tilde{I}_i \, A_i(\xi_i) \tag{7}$$

$$\tilde{I}_i = -\frac{1}{2}\log(1-\rho_{0,i}^2) \tag{8}$$

$$A_i(\xi_i) = \frac{\log\frac{(1-\rho_{0,i}^2)\xi_i+1}{\xi_i+1}}{\log(1-\rho_{0,i}^2)}. \tag{9}$$

The maximum mutual information is attained at high SNR and is $\sum_i \tilde{I}_i$. Defining $I_i = \tilde{I}_i / \sum_j \tilde{I}_j$ and normalizing (7) accordingly, we recognize I_i as the so-called *band-importance function* and $A_i(\xi_i)$ as the so-called *weighting function* or *band-audibility function*. The formulation (7) forms the basis of speech intelligibility measures such as the SII [18] and the Extended SII [27]. These measures are descendants of the so-called *articulation index* [24], [25], a measure that predates information theory. In this classic view, I_i characterizes the importance of frequency band i and the factor A_i is a weighting function that indicates what fraction of the information is delivered to the listener. The information-theory derived form of A_i shown in (9) describes a sigmoidal function that approximates the definition of A_i in the SII. ((9) neglects the threshold of hearing, the effect of high loudness, and the self-masking of noise.) Our derivation of the band importance function I_i of (8) makes its dependency on the production and interpretation noise explicit. If the relative variances of the production and interpretation noise of a band are low (high production and interpretation SNR; $\rho_{0,i}$ approaches one), that band is important for intelligibility. In the SII definition, the values of I_i are set empirically. As is shown in [15], the differences between the formulas for the classic approach and the above information-theoretical derivation are well within the precision of the original heuristic derivation of the classic view. The classic SII has proven to be highly correlated with speech intelligibility in many conditions and has been used as a basis for speech intelligibility enhancement [4], [8], [12], [28]. It is discussed in additional detail in Section III-B1.

4) Constraints on Optimization: In most cases, the optimization must be performed subject to one or more constraints. Important constraints are the speech-like nature of the output, the signal power, and system delay. Additional constraints may be required. For instance, for a given message M_T (and speaking rate), a longer word sequence will likely be more intelligible than a short one, thus making a length constraint natural.

The speech-like nature, or the speech quality, of the enhanced output may require an explicit constraint. However, in most practical systems the speech-like nature is enforced implicitly by either the modification strategy, or the optimization criterion, or both. Modification strategies such as slowly varying spectral shaping facilitate speech-like output only. The maximum probability of correct phoneme recognition is an example of a criterion that favors signal features that that resemble those of clean speech.

Signal power is a natural constraint. The unconstrained optimization of signal spectral modifications may lead to an unbounded increase of the signal power if the reduction in recognition performance of the human auditory system for loud sounds is not considered. Thus, a power constraint must be applied to prevent hearing injuries and loudspeaker damage. Approximations to perceived loudness, either in the form of an analytic expression, or in the form of an algorithm, may also be used as constraints.

The system delay must be constrained in real-time systems. This may prevent the usage of particular distortion measures and modification operators.

III. Signal Processing Approaches

In this section, the focus is on creating practical enhancement systems. We start with a discussion of various modifications that can be made and then discuss three approaches to enhancement and their performance. The boxes describe specific applications.

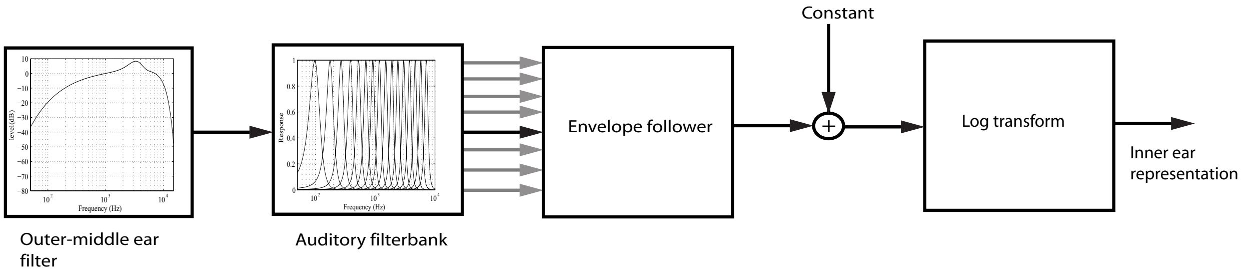

Fig. 1: Basic structure of the auditory model presented in [21].

A. Speech Modifications

The basic paradigm of intelligibility enhancement discussed in this paper is to select a modification operation to be used for preprocessing the signal and a measure of intelligibility, and then to adjust the parameters of the modification operation to maximize the measure. Below, we discuss the classes of modifications that have been used or can be used and report on current knowledge about their effectiveness.

Enhancement operators can be classified according to a number of criteria. Operators can be classified generically as time-varying or time-invariant and as linear or non-linear. Most intelligibility enhancement operators are time-invariant and nonlinear. However, low-level operators that use a linear filtering of the signal [8] have been used and perform well (if the filter is adapted, the operator is nonlinear).

Additional classifications can be made based on the specific processing performed on the message. Depending on the abstraction level where a modification takes place, we identify lexical (high-level), prosodic (mid-level) and spectral and temporal (low level) modifications. In accordance with the Markov chain model of the communication process, presented in Section II-A, a high-level modification affects the message representation at the lower levels. The operator can be independent or dependent on the environmental disturbance, i.e., it can be non-adaptive or adaptive. Finally, depending on the origin of a modification there are i) mimicking strategies, *i.e.*, modifications that attempt to mimick modifications used consciously or subconsciously by humans producing speech in adverse conditions, and ii) rational strategies based on, *e.g.*, expert insight in the human auditory periphery and in cognition [3] or of the sound field [16], [29].

In unpublished work of the Listening Talker (LISTA) project
(http://listeningtalker.org) 44 possible modifications were identified. This includes the modification strate-

gies used in essentially all existing intelligibility enhancement systems. The effectiveness of some of the listed modifications on the intelligibility in noisy environments is reviewed in [7], [9].

Making mobile phones more intelligible: Mobile telephony is often conducted in the presence of acoustical background noise such as traffic or babble noise. In this situation, the listener perceives a mixture of clean speech and environmental noise from the near-end side, which generally leads to an increased listening effort and possibly to reduced speech intelligibility. As the noise signal generally cannot be changed, the manipulation of the far-end signal is the only way to effectively improve speech intelligibility and to ease listening effort for the near-end listener.

In the mobile-phone application, the algorithmic delay of the processing is crucial since the allowed roundtrip delay of the communication system is limited. This places a severe constraint on the modification operator. Furthermore, the restrictions of the micro-loudspeakers of mobile phones need to be considered. The maximum thermal load of the micro-loudspeaker constitutes a major limitation, which can be taken into account with a constraint on the total audio power. Finally, the ear of the near-end listener is usually next to the loudspeaker and must be protected from damage and pain. This can be ensured by a power limitations for the critical bands.

As is discussed in [7], [9], mimicking strategies such as pitch modification, vowel space adjustment and uniform speaking rate reduction do not improve intelligibility consistently when applied to natural speech. This outcome suggests that such modifications may have an auxiliary role or may be the result of physical limitations in the speech production mechanism. Other mimicking candidate modifications include changing the relative duration of phonetic units and shortening units that are more sensitive to energetic masking in favor of more robust units. As yet no conclusions can be drawn about the benefit from such modifications. In the remainder of this section we focus on rational strategies.

Lexical speech modifications consist of, among others: i) repetition to provide additional cues and ii) rephrasing to increase correct recognition probability as a result of better noise robustness and/or higher predictability. While repetition does not facilitate intelligibility optimization, rephrasing provides an intuitive and attractive modification class. Section II discussed high-level modification measures that can, at least in principle, be used for this purpose. A practical rephrasing approach is presented in [19]: rather than comparing the measures directly, the method compares the sensitivity to noise addition of each formulation, according to the probability of correct recognition. The approach does not consider the predictability of the formulation, which is a major factor in intelligibility. An indirect indication of the expected gain from increasing the predictability of a formulation, *e.g.*, by vocabulary size reduction, can be obtained by comparing the outcomes of intelligibility evaluations using closed-set [14] and open-set vocabulary bases [9]. The considerably higher intelligibility gain for closed-set evaluation suggests that it is feasible to design a modification system achieving intelligibility gain by improving the predictability of the formulation.

Making it work for hearing instruments:

Hearing instruments aim to compensate for a hearing loss. Typically, this is done by amplifying a sound recording, followed by dynamic range compression to ensure the signal remains within the audible and comfortable range. Environmental noise degrades intelligibility for hearing instrument users in two ways. A first degradation is due to noise recorded by the microphones. To decrease the impact of this noise, noise reduction is applied to the recorded signal prior to amplification for hearing loss compensation.

A second degradation depends on the fitting: the user may experience direct environmental noise, leaking through the hearing instrument vent. This leakage degrades the intelligibility and can be overcome by processing the signal with the application of a speech intelligibility enhancement algorithm before playout as discussed in the main text.

Adopting the concept of interpretation noise, the patient's hearing loss can be measured and modeled by the noise process $\mathbf{v}_L$. The environmental noise that reaches the ear through the hearing instrument vent can be modeled by the process $\mathbf{v}_E$ of (4). Dynamic range compression can be taken into account by expressing the desired output range in terms of (frequency dependent) absolute power constraints. Given this model, the hearing instrument can be optimized using one of the measures discussed in Sec. II-B3 in a constrained fashion. The resulting integrated solution compares favorably with an ad-hoc concatenation of processing steps, facilitates a conceptual understanding of the hearing impairment and is likely to lead to an effective control of the instrument.

Low-level modifications do not require knowledge of the intended message transcription. These can be subdivided into spectral, temporal, and spatial signal modifications as well as combinations there-of.

Straightforward spectral shaping is employed in [8], [12]. This modification facilitates both low complexity and a high intelligibility gain, *e.g.*, [9] making these approaches particularly suitable for application in mobile telephony.

Spectro-temporal energy redistribution is considered in [6] where the glimpse proportion is optimized. Use of a genetic algorithm to perform the optimization makes this method interesting primarily from a theoretical perspective. A low-complexity approach with high intelligibility gain that performs spectro-temporal energy redistribution by optimizing a perceptual distortion measure is presented in [14].

A particular class of spectro-temporal energy redistribution is obtained with dynamic range compression. This approach can either be non-adaptive or adaptive. In a large-scale subjective evaluation of proposed speech modification systems [9], most of the entries that incorporated dynamic range compression, including those related to the descriptions in [5], [23], [28], performed well.

Intelligibility can also be enhanced by controlling the spatial soundfield near the ear with a multitude of remote loudspeakers. As discussed in more detail in section III-B3, if users are wearing microphones near their ears, reverberation and cross-talk between different messages can be reduced by feedback [16]. The goal is that only the desired signal is present at the ear of a user. If microphones are further from the ears of the listeners, the emerging field of multi-zone audio rendering becomes relevant, *e.g.*, [29].

B. Intelligibility Enhancement Systems

This section describes three practical methods for intelligibility optimization approaches. The described approaches are based on different principles.

1) SII based Enhancement: State-of-the-art systems have been developed based on the decomposition into band-importance and band-audibility functions [4], [8], [28]. We provided a recent perspective on this decomposition in Section II-B3. The present section describes implementations that closely follow the SII standard.

The computation of the SII [18] uses a carefully calibrated specification of the speech spectrum $\sigma^2_{\mathbf{a}_{T,i}}$ and the noise spectrum $\sigma^2_{\mathbf{v}_{E,i}}$ (where i is a critical or third-octave band index) as measured over an entire utterance, including minor pauses. The approach accounts for both the hearing threshold and the loss of intelligibility at very high presentation (loudness) levels, using information stored in tables. For an acoustic time-domain speech signal $\mathbf{a}_T$, the *equivalent speech spectrum level* in dB, commonly denoted as E_i, is computed as

$$E_i = 10\log_{10}(\frac{\sigma^2_{\mathbf{a}_{T,i}}}{f_{\Delta,i}}) - 10\log_{10}(\sigma_0^2), \tag{10}$$

where σ_0^2 denotes the digital reference power per Hz corresponding to the reference sound pressure of 20 µPa and $f_{\Delta,i}$ is the frequency bandwidth of the i-th subband in Hz. The *equivalent disturbance spectrum level*, D_i, is computed in three steps: first the calibration (10) is applied, and then the threshold of hearing and instantaneous masking are accounted for. In [4] the threshold of hearing and in [8] both the threshold of hearing and instantaneous masking are neglected.

The band-audibility function of the SII also accounts for the decrease in intelligibility at high presentation (loudness) levels, which is not accounted for in (9). Consequently, it depends on both the SNR in the band and the absolute presentation level E_i. The band-audibility function is identical for different bands and we denote it for a band i as $A(E_i, D_i)$. Let us define the piecewise linear sigmoid $\mathcal{S}(x; \beta_1, \beta_2) = (\max(\min(x, \beta_2), \beta_1) - \beta_1) / (\beta_2 - \beta_1)$, which has a range $[0, 1]$. The band audibility function of the SII is factorized into two factors: the first factor accounts for the instantaneous masking and the second factor accounts for high presentation levels:

$$A(E_i, D_i) = \mathcal{S}(E_i; D_i - 15, D_i + 15)\, \mathcal{S}(-E_i; -U_i - 170, -U_i - 10), \tag{11}$$

where U_i is the standard speech level at normal voicing effort (provided in a table in the standard). The heuristic factor $\mathcal{S}(E_i; D_i - 15, D_i + 15)$ assumes that speech signals 15 dB below the disturbance level are fully masked, and speech signals 15 dB above the disturbance level are not masked, which leads to a curve similar to the result derived in (9).

The SII is a refined and normalized version of (7) that accounts for decreased intelligibility at high presentation levels:

$$\mathrm{SII} = \sum_i I_i\, A(E_i, D_i)\,. \tag{12}$$

The band-importance function I_i in the SII is specified by a table that is based on fitting to a database. Fig. 2 illustrates the computation of the SII. The suppression of the audibility function at high presentation levels is clearly shown in the panel showing the audibility function (11).

The measure (12) can be used to optimize a modification operator that shapes the spectrum. As the intelligibility decreases both at high and low presentation levels, the SII criterion can, in principle, be optimized without constraint. It is seen from (12) that if there is no global constraint, each frequency band can be optimized independently. The resulting solutions are not necessarily unique because of the form of $\mathcal{S}$. It is natural to select the solution that has the lowest power, but does not reduce the speech power in any band. For low absolute noise levels, where the solution is not limited by the second factor in (11), the solution for the gain is [4]

$$g_i = \max\left(D_i + 15, E_i\right) - E_i, \tag{13}$$

where the shaping gain g_i for band i is given in dB. In (13) the original equivalent speech spectrum level is E_i and the modified speech has equivalent speech spectrum level $g_i + E_i$.

As was discussed in section II-B4, it is common to constrain the overall loudspeaker signal power in practical applications. The optimization of (12) subject to a power constraint was studied in [4] and [8]. To facilitate analysis, the two approaches use approximations of (12). Althought the approximations are different, both neglect the second factor in (11) and start from $A(E_i, D_i) \approx \mathcal{S}(E_i; D_i - 15, D_i + 15)$. [4] simplifies $A(E_i, D_i)$ further by removing the lower bound on the sigmoid and writing $A(E_i, D_i) \approx \frac{1}{2} +$

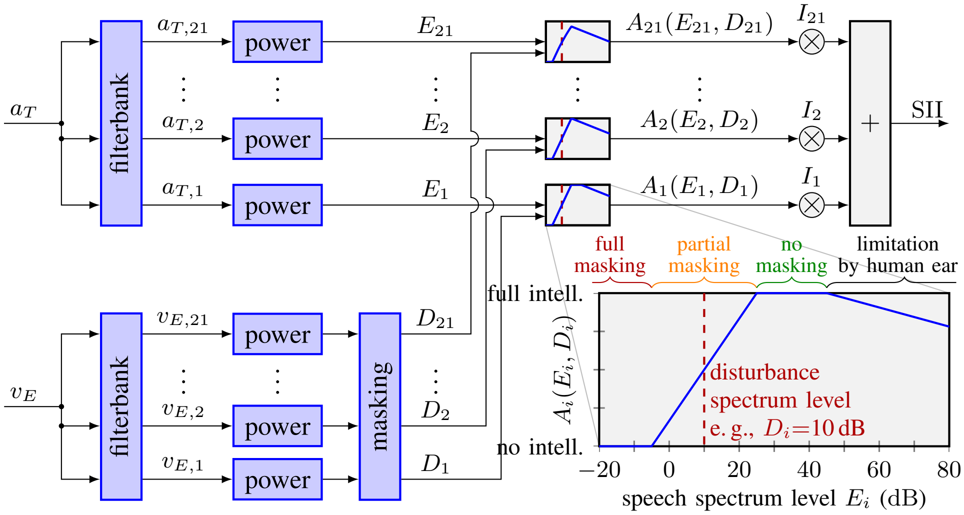

Fig. 2: The computation of the SII.

$\min(E_i - D_i, 15)/30$. [8], on the other hand, makes the approximation $A(E_i, D_i) \approx 10^{E_i/10}/(10^{E_i/10} + 10^{D_i/10})$, which is a differentiable function. When writing the above expressions for the modified speech, the audibility-function approximations are concave functions of the (linear) spectral gain $10^{g_i/10}$. Optimizing the approximations subject to linear constraints on $10^{g_i/10}$ form straightforward optimization problems that can be solved using the Karush-Kuhn-Tucker conditions. The resulting analytic solutions are easy to implement. The later work of [8] models $A(E_i, D_i)$ more accurately at low SNR values and provides improved performance over the original work of of [4] under low SNR conditions.

The discussion in this section assumed stationarity. Time variation can be accounted for by recursive updating of the equivalent spectrum levels E_i and D_i and periodically recomputing the gains g_i [4]. This is consistent with the SII update described in [27], which uses frequency-dependent temporal windows.

2) Word-sequence Probability based Enhancement: Section II-A identified the suitability of the expected probability of correct message recognition as a measure for optimizing intelligibility at a high level of abstraction. We noted in Section II-B1 that, under an ergodicity assumption, the expectation over messages can be approximated by averaging over time. Optimizing a measure derived from the probability of correct recognition under a power constraint has been shown to provide significant intelligibility gain assuming that accurate sound segmentation information and an appropriate acoustic speech model are available [11]. We emphasize that the method assumes that ASR word probability tracks the human recognition performance, which was found to be true in [11] but is not guaranteed. In this subsection we provide more detail about this approach.

To make high-level machine-based optimization feasible in practice, we can represent the message at the phoneme level. This means we refine our Markov chain to include an intermediate level. The chain now becomes $\mathbf{M}_T \rightarrow \mathbf{u}_T \rightarrow \mathbf{a}_T \rightarrow \mathbf{a}_L \rightarrow \mathbf{u}_L \rightarrow \mathbf{M}_L$, where $\mathbf{u}_T$ and $\mathbf{u}_L$ denote the talker and lister phoneme sequences, respectively. By first performing time alignment of a sequence of acoustic features vectors $\mathbf{a}_T$ and a sequence of phonemes $\mathbf{u}_T$ by means of an ASR engine, a practical intelligibility enhancement approach can be defined. The ASR speech model can then be used to provide the probability densities that characterize clean speech sounds in the acoustic feature space.

To enhance intelligibility, we want to find the parameters $\mathcal{C}^*$ of our speech modification scheme that maximize the average probability that the listener interpreted phoneme sequence $\mathbf{u}_L$ is the talker-generated sequence $\mathbf{u}_T$:

$$\mathcal{C}^* = \underset{\mathcal{C}}{\operatorname{argmax}} \; \overline{p_{\mathbf{u}_L|\mathbf{u}_T}(u_T|u_T, \mathcal{C})} \tag{14}$$

where the subscripts of the density label the density it represents. Note that the densities are consistent with the models shown in (4).

Simplifications were introduced in [11] to make the optimization tractable. It was tacitly assumed that the message is accurately represented by the phonemes and production noise was not formally considered. It was also assumed that v_E (the representation of the noise) can be approximated as deterministic, which is reasonable for typical acoustic signal representations and stationary noise. The only remaining uncertainty is due to the interpretation noise in the mapping from a_L to u_L. In an ASR system based on an HMM this is modeled by the observation noise. Eq. (14) can now be approximated by:

$$\mathcal{C}^* \approx \underset{\mathcal{C}}{\operatorname{argmax}} \; \overline{p_{\mathbf{u}_L|\hat{\mathbf{a}}_L}(u_T|\hat{a}_L(u_T, \mathcal{C}))}, \tag{15}$$

$$= \underset{\mathcal{C}}{\operatorname{argmax}} \; \overline{p_{\hat{\mathbf{a}}_L|\mathbf{u}_L}(\hat{a}_L|u_T, \mathcal{C})\; p_{\mathbf{u}_L}(u_T) \; \left(\sum_{u'_T} p_{\hat{\mathbf{a}}_L|\mathbf{u}_L}(\hat{a}_L|u'_T, \mathcal{C}) p_{\mathbf{u}_L}(u'_T)\right)^{-1}}, \tag{16}$$

where we used Bayes' rule and where $\hat{a}_L(\mathbf{u}_T, \mathcal{C})$, abbreviated to $\hat{\mathbf{a}}_L$, is the set of acoustic features observed by the listener, which is modeled as a deterministic function of the talker phoneme sequence u_T and the speech modification parameters $\mathcal{C}$. The first term of (16) is the likelihood of the talker phoneme sequence for the observed features $\hat{\mathbf{a}}_L$, the second term is the *a-priori* probability that the phoneme sequence $\mathbf{u}_T$ is decoded by the listener, and the third term is the inverse *a-priori* probability of the listener observed features. Optimization of the likelihood term only reduces complexity and provides good results [11].

The theory is simplest to implement if the sequences are considered stationary. Averaging of (16) over long time intervals (multiple sentences) is then preferred. In a practical implementation, shortcuts may

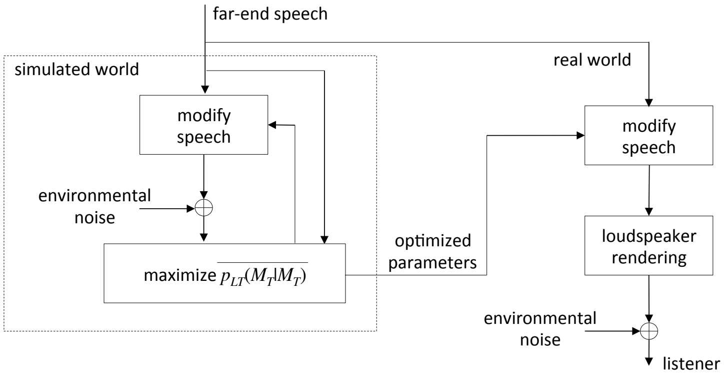

Fig. 3: Intelligibility enhancement using a phoneme-level measure.

have to be made due to requirements on delay and complexity and because the stationarity assumption may not be sufficiently accurate.

A system-level perspective of the proposed approach is shown in Figure 3. In [11], the approach was validated for a combination of two modifications: prosody-affecting phoneme gain adjustment and a spectral modification redistributing the signal energy across frequency bands. The method compared favorably to a method based on the optimization of a measure operating on a sequence of auditory states [14], discussed in Sec. II-B2. Results reported in [9] suggest that using the full Bayesian approach rather than optimizing only the likelihood component of (16) improves performance.

In text-to-speech applications it may be possible to select from a set of phrases to convey a particular message. The measure given in (16) has also been used to determine the optimal phrasing of utterances [19]. This study indicates that maximizing the probability of correct interpretation of the phoneme sequence increases intelligibility. Taking into consideration prior information on the predictability of the various formulations is expected to enhance performance further.

3) Enhancement over Multiple Spatial Points: Thus-far we have considered preprocessing techniques that do not consider the spatial aspects of the rendering scenario. In this section we show that spatial aspects can also be exploited to enhance intelligibility. In announcement scenarios in public spaces such as airports, train stations or shopping malls, environmental noise and reverberation contribute to a reduced intelligibility at the listeners. If different messages are communicated to different spatial regions, acoustic leakage between regions [16] exacerbates the problem. The impact on intelligibility is particularly large for hearing-impaired persons.

Consider a scenario in a public environment where N messages are conveyed via the public address

(PA) system to N listeners wearing a hearing instrument. A possibility is to downstream the corresponding signals directly to the listeners, but listeners often wear an open fit (non-occluded) hearing instrument, where the direct signal also is mixed in at the eardrum. Instead of using direct downlink connections, it is possible to pre-process all speech signals jointly at the PA system so as to minimize the expected distortion at the eardrums of the listeners. The distortion measure can be based on any (mathematically well-behaved) model for speech quality or intelligibility, such as some of the models discussed in Sec. II-B.

Let $a_T = [a_{T,1}, a_{T,2}, \ldots, a_{T,N}]^{\mathrm{T}}$, $\tilde{a}_T$ and $\boldsymbol{a}_L$ (defined similarly) be the (complex-valued) short-time DFT coefficients of the source speech signals, enhanced signals (at the PA system), and received signals at the listeners, respectively. The signals $\boldsymbol{a}_L$ are captured by the microphones of the hearing instruments. For simplicity, we neglect production and interpretation noises of Sec. II-A and assume that degradations are purely acoustical and consist of noise, reverberation and crosstalk between messages. It is easy to see that if we use stacked-vector notation for the signals $a_{T,i}$ and $\boldsymbol{a}_{L,i}$, $i = 1, 2, \ldots, N$, upon pre-processing, all effects can be included in the affine signal model given by [16]

$$\boldsymbol{a}_L = \boldsymbol{H}_E \tilde{a}_T + \boldsymbol{v}_E \, , \tag{17}$$

where the channel matrix $\boldsymbol{H}_E$ collects all reverberation and crosstalk transfer coefficients between production and reception points, and $\boldsymbol{v}_E$ is additive noise in the environment.

Consider also a distortion measure $d(a_T, a_L)$, smooth (continuously differentiable) as a function of a_L, which quantifies the distortion between the reference produced coefficients a_T and what is eventually listened to, a_L. Our aim is to find the modification $a_T \mapsto \tilde{a}_T$ that minimizes the expected distortion according to d, jointly for all talker-listener points, *i.e.*, we want to solve the optimization problem

$$\underset{\tilde{a}_T}{\text{minimize}} \;\; \mathrm{E}[d(a_T, \boldsymbol{H}_E \tilde{a}_T + \boldsymbol{v}_E)] \, , \tag{18}$$

where the expectation is taken only over the acoustic disturbances $\boldsymbol{H}_E$, $\boldsymbol{v}_E$, since we have direct access to the speech of the talker a_T and therefore take it to be deterministic.

Generic necessary conditions can be derived for solving (18) in terms of a functional description of the distortion measure d. The conditions are [16]

$$\mathrm{E}\left[\boldsymbol{H}_E^{\mathrm{H}} \frac{\partial d}{\partial a_L^*} (a_T, \boldsymbol{H}_E \tilde{a}_T + \boldsymbol{v}_E) \right] = 0 \, , \tag{19}$$

where $(\cdot)^{\mathrm{H}}$ is the hermitian transpose, and $\frac{\partial}{\partial v^*} \equiv \frac{1}{2} \left(\frac{\partial}{\partial v_{\Re}} - \frac{1}{j} \frac{\partial}{\partial v_{\Im}} \right)$ is a complex differential operator, expressed in terms of the real differential operators $\frac{\partial}{\partial v_{\Re}}$ and $\frac{\partial}{\partial v_{\Im}}$, in Hessian (vertical) notation, with respect to the real and imaginary components of the variable v, respectively. The meaning of (19) is that, for optimality, it is required to choose the pre-processed speech $\tilde{a}_T$ such as to make the complex gradient of the distortion measure with respect to the listener DFT bins in all zones orthogonal to all columns of the channel matrix $\boldsymbol{H}_E$.

To demonstrate the use of the optimality conditions (19), let us consider the simple ℓ_2 distortion measure given by

$$d(a_T, a_L) = \|a_L - a_T\|^2 \,, \tag{20}$$

where $\|\cdot\|$ is the ℓ_2 norm. In this case, (18) is a convex optimization problem, so that (19) are also sufficient conditions. By using the optimality conditions (19) under the assumption that $\boldsymbol{H}_E$ and $\boldsymbol{v}_E$ are uncorrelated, and including the hybrid deterministic-stochastic model for $\boldsymbol{H}_E$ introduced in [16], where the early response is described solely by a deterministic direct path and the late response is modeled by an exponentially fading stochastic process, the pre-processing algorithm is derived as

$$\tilde{a}_T = \left(D^{\mathrm{H}} D + \Lambda\right)^{-1} D^{\mathrm{H}} a_T \,, \tag{21}$$

where D is a matrix collecting direct path responses of the channel, and Λ is a diagonal matrix collecting diffuse reverberation response channel energies. Note that in the case of low reverberation, $\Lambda \to 0$, the scheme (21) reduces to a conventional acoustic crosstalk canceller [30], $\tilde{a}_T = D^{-1} a_T$, which by compensating for the direct paths of the channel $\boldsymbol{H}_E$, makes the cross-signals cancel out at the listeners. We thus conclude that optimization-based multipoint pre-processing enhancement as formulated in (18) leads to acoustic crosstalk cancellation, when applied to the ℓ_2 distortion measure (20).

IV. Conclusions / Open Problems

Modern speech communication often leads to the signal being rendered by a machine in a noisy environment. In these circumstances, communication benefits from methods that make speech more intelligible in noise, particularly if the enhancement can adapt to the scenario at hand. This requires quantitative models of the communication process and distortion measures.

The use of a distortion measure facilitates the formulation of convergent algorithms and generally reduces the need for ad-hoc solutions. Measures formulated at a high level of abstraction, such as (1) and (3) apply, at least in principle, to all communication tasks. However, when these high-level measures are applied to specific tasks assumptions must be made, either for the signal or for a model of the human cognitive system (e.g., by an ASR system), or both. Thus, optimization of any measure can never replace the need of extensive real-world testing to verify the performance of an intelligibility-enhancement system for the task at hand.

At first sight the intelligibility-enhancement problem resembles the standard problem of transmission over a noisy channel. However, we have shown that the unprecise nature of the human production and interpretation must be accounted for. When that is done, standardized measures for intelligibility, which have a long history and were derived heuristically, are found to be consistent with communication theory.

While the field of intelligibility enhancement has developed rapidly, opportunities for significant improvement remain. Careful accounting for time-domain masking may improve performance. Methods developed for scenarios with additive noise only must be extended to include reverberation. Refining methods that perform spectral shaping to include range compression may increase their performance. For methods based on mutual information, the effect of time and frequency dependencies must be considered. Studies to determine the best representation (e.g., cepstra or DFT coefficients) and the determination and usage of appropriate noise distributions for the model likely will lead to improvement. The determination of a word choice for a message that is more robust to noise is an essentially unsolved task.

Although major challenges remain, the field of intelligibility enhancement has made major strides in recent years. The technical outcomes will likely become an integral part of speech-rendering devices in the near future, leading to improved communication among humans and from machines to humans.

References

[1] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," J. Acoust. Soc. Am., vol. 128, pp. 2059–2069, 2010.

[2] J. D. Griffiths, "Optimum linear filter for speech transmission," J. Acoust. Soc. Am., vol. 43, no. 1, pp. 81–86, 1968.

[3] 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. 24, no. 4, pp. 277–282, 1976.

[4] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in ITG-Fachbericht-Sprachkommunikation, 2010.

[5] 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, USA, 2012.

[6] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in ISCA Interspeech, Portland, USA, 2012.

[7] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," Elsevier Comput. Speech Language, vol. 28, pp. 543–571, 2014.

[8] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," Signal Processing Letters, IEEE, vol. 20, no. 3, pp. 225–228, 2013.

[9] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in ISCA Interspeech, Lyon, France, 2013.

[10] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," Speech Communication, vol. 55, pp. 572–585, 2013.

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

[12] B. Sauert and P. Vary, “Near end listening enhancement optimized with respect to speech intelligibility index,” in EURASIP Europ. Signal Process. Conf. (EUSIPCO), vol. 17, 2009, pp. 1844–1848.

[13] M. Cooke, “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1562–1573, 2006.

[14] C. H. Taal, R. C. Hendriks, and R. Heusdens, “Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure,” Computer Speech & Language, 2014.

[15] W. B. Kleijn and R. C. Hendriks, “A simple model of speech communication and its application to intelligibility enhancement,” Signal Processing Letters, IEEE, vol. 22, no. 3, pp. 303–307, March 2015.

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

[17] J. Allen, “How do humans process and recognize speech?” Speech and Audio Processing, IEEE Transactions on, vol. 2, no. 4, pp. 567–577, Oct 1994.

[18] ANSI S3.5-1997, “Methods for the calculation of the speech intelligibility index,” Am. National Standards Inst., 1997.

[19] M. Zhang, P. N. Petkov, and W. B. Kleijn, “Rephrasing-based speech intelligibility enhancement,” in ISCA Interspeech, Aug. 2013.

[20] 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. Am., vol. 99, pp. 3615–3622, 1996.

[21] C. H. Taal, R. C. Hendriks, and R. Heusdens, “A low-complexity spectro-temporal distortion measure for audio processing applications,” IEEE Trans. on Audio, Speech, and Language Process., vol. 20, no. 5, pp. 1553–1564, 2012.

[22] C. Valentini-Botinhao, J. Yamagishi, and S. King, “Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?” in ISCA Interspeech, August 2011.

[23] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, “Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the Glimpse Proportion,” Computer Speech and Language (in press), 2013.

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

[25] K. D. Kryter, “Methods for the calculation and use of the Articulation Index,” J. Acoust. Soc. Am., vol. 34, no. 11, pp. 1689–1697, November 1962.

[26] G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, “A frequency importance function for continuous discourse,” J. Acoust. Soc. Am., vol. 81, no. 4, pp. 1130–1138, 1987.

[27] 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. Am., vol. 117, no. 4, pp. 2181–2192, 2005.

[28] H. Schepker, J. Rennies, and S. Doclo, “Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression,” in ISCA Interspeech, 2013.

[29] S. Elliott, J. Cheer, J.-W. Choi, and Y.Kim, “Robustness and regularization of personal audio systems,” IEEE Trans. Speech, Audio and Language Process., vol. 20, pp. 2123–2133, Sept. 2012.

[30] D. B. Ward and G. W. Elko, “Virtual sound using loudspeakers: robust acoustic crosstalk cancellation,” in Acoust. Signal Proc. for Telecom, S. L. Gay and J. Benesty, Eds. Boston, MA: Kluwer Academic, 2000, ch. 14.