Maximizing Phoneme Recognition Accuracy for Enhanced Speech Intelligibility in Noise

Petko N. Petkov, *Student Member, IEEE,* Gustav Eje Henter, *Student Member, IEEE,* and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—An effective measure of speech intelligibility is the probability of correct recognition of the transmitted message. We propose a speech pre-enhancement method based on matching the recognized text to the text of the original message. The selected criterion is accurately approximated by the probability of the correct transcription given an estimate of the noisy speech features. In the presence of environment noise, and with a decrease in the signal-to-noise ratio, speech intelligibility declines. We implement a speech pre-enhancement system that optimizes the proposed criterion for the parameters of two distinct speech modification strategies under an energy-preservation constraint. The proposed method requires prior knowledge in the form of a transcription of the transmitted message and acoustic speech models from an automatic speech recognition system. Performance results from an open-set subjective intelligibility test indicate a significant improvement over natural speech and a reference system that optimizes a perceptual-distortion-based objective intelligibility measure. The computational complexity of the approach permits use in on-line applications.**

***Index Terms*—environment adaptation, intelligibility enhancement, speech pre-enhancement**

I. Introduction

Intelligibility is a quantitative representation of the similarity between the perceived and the intended messages. The search for modification strategies improving speech intelligibility in noisy environments is an active research topic. While the focus is primarily on additive distortions [1], [2], [3], application to other impairments, such as reverberation, have been considered as well [4]. In this work we consider additive distortions because of their broad practical significance.

A source of inspiration in the search for speech modifications are the strategies adopted by talkers to counteract the effect of noise on intelligibility. The Lombard effect reveals that apart from loudness, changes to the speaking rate, spectral tilt and the pitch are also produced [5], [6], [7]. While changes in some speech parameters, such as pitch, do not improve intelligibility on their own and are likely the effect of constraints on the speech production mechanism, Lombard speech is more intelligible than natural speech in noise, even when the loudness of the two is equalized. Other human strategies such as vocabulary adjustment [8], emphasis of information-bearing word types [9] and repetition can prove effective as well.

From the perspective of automated speech modification, a number of methods for the enhancement of speech intelligibility in noisy environments have been proposed. These can be classified into two main categories: i) rule-based, e.g., [1], [10], [2], [11], [12], [13] and ii) objective-intelligibility-measure-based [14], [15], [16], [3]. Rule-based methods operate along a framework of heuristic rules for modification that have been identified to produce intelligibility improvement in given noise contexts. The use of objective intelligibility measures such as the speech intelligibility index (SII) [17], the glimpse proportion (GP) [18] and perceptual distortion (PD) [3], identifies a more recent stage in the evolution of the field. Methods from this category optimize an objective measure, which approximates intelligibility, for the parameters of a speech modification. A reliable measure can, in theory, be used with a range of modification strategies, establishing the measure-based approach as the more general one.

The interaction between speech modification and measure optimization and the implications of a particular choice for a modification strategy and an objective measure can be illustrated in the presence of a model of the speech communication process. Inspired by the hierarchical model of human speech recognition of [19], and its representation from [20], we associate the communication process with a first-order Markov chain as illustrated in Figure 1. Each state of the depicted one-way communication chain, is conditionally-independent of previous states given the preceding state. The speaker formulates a message based on the need to convey certain information to the listener. The message is then translated into a sequence of phonemes reflecting the language proficiency of the speaker. The voice production mechanism is activated and the vocal tract articulators [21] manipulated such that the desired sound can be produced. Co-articulation [22] (not illustrated) introduces a dependence of individual sounds on their phonetic context.

A speaker adjusts the speech production process to adapt to environment noise and increase the probability for correct decoding of the message. Some adjustments are listed to the left of the speaker-side part of the communication Markov chain in Figure 1. These are aligned with the chain state at which they operate. *Formulation* includes vocabulary and sentence structure simplification, and repetition of salient

P. N. Petkov is with the School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm 100 44, Sweden (e-mail: petkov@ee.kth.se).

G. E. Henter is with the School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm 100 44, Sweden (e-mail: gustav.henter@ee.kth.se).

W. B. Kleijn is with the School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm 100 44, Sweden and the School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand (e-mail: bastiaan.kleijn@ecs.vuw.ac.nz).

segments. *Pronunciation* takes into account, e.g., on-set and transient amplification, and accent adaptation. *Coloring* stands for various spectral modification not related to the two previous categories.

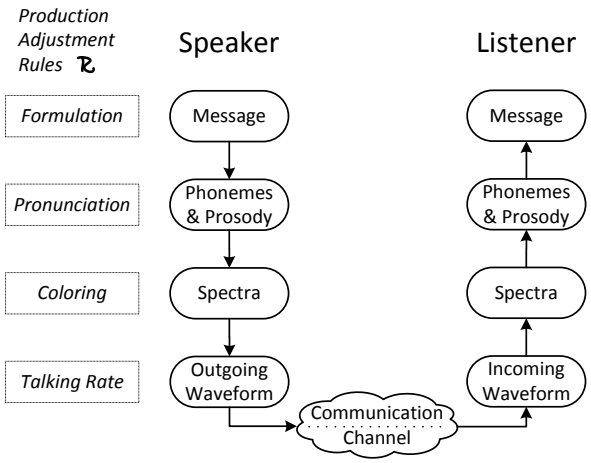

Fig. 1. The communication process represented as a first-order Markov chain. Thus, e.g., the outgoing waveform is conditionally-independent of the message, the phoneme sequence and signal prosody given the spectrum.

An objective measure approximating intelligibility can be formulated at any of the listener-side communication hierarchy levels below that of the message. At the *Phonemes & Prosody* level intelligibility can be approximated by means of, e.g., automatic speech recognition (ASR) [18]. At the *Spectra* and *Waveform* levels objective measures commonly exploit knowledge of the band-specific noise power spectra and signal-to-noise ratios [17], [18], [23]. Current methods for speech pre-enhancement, such as [14], [3], [15], [16], optimize objective intelligibility measures at the lower hierarchy levels.

The choice between a low and a high level objective intelligibility measure is guided by the trade-off between efficiency and optimality. If we control the result of the optimization at a level closer to that of subjective intelligibility (the message level), the possibility for divergence between the objective implied by the measure and the true objective of improving subjective intelligibility is reduced. Furthermore, for measures higher in the hierarchy, the number of visible modification strategies increases. Consider, e.g., prosody modification and vocabulary adjustment that remain transparent to a measure such as SII but can effectively increase intelligibility. The challenges with using high-level measures are related to i) the need for an exact or an estimated transcription of the message, and ii) the increasing complexity. The second aspect is clearly illustrated by comparing the complexity of SII [17] and PD [3] to that of the glimpsing model [18].

In this paper, we show that a practical objective intelligibility measure for application to speech pre-enhancement can be defined at the phoneme level. To achieve this we employ a word-level transcription of the transmitted message and a statistical model of speech from ASR. The method is therefore best-suited for application in text-to-speech and recorded speech modification. An adaptation to live speech, however, can be considered as well. The proposed measure is sufficiently general to accommodate a range of modification strategies including spectral and temporal domain energy reallocation. An early version of the method is described in [24]. The results from a pilot experiment were presented in [25].

Advances in ASR gradually close the gap from human speech recognition [20]. The use of a conventional speech model from ASR in our experiments suggests that a perfect recognition model is not needed to achieve intelligibility enhancement. Subjective evaluation of the proposed approach in a large-vocabulary, open-set test demonstrated significant improvement over natural speech in noise for both speech-shaped and babble noise. The proposed method significantly outperforms the reference method [3] in low SNR conditions. The reference method has similar modification capabilities but does not assume knowledge of the message transcription or access to a speech model from ASR. The proposed method is more complex than its counterpart. By circumventing the need for a complete recognition system, however, the complexity is kept at a level feasible to accommodate in on-line applications.

The remainder of this paper is organized as follows. The theoretical foundations for our work are established in Section II. The speech pre-enhancement system is described in Section III. The experimental set-up and the validation results are discussed in Section IV. Section V presents the conclusions.

II. Theoretical Foundations

We present the derivation and the analysis of the proposed measure of objective intelligibility at the text level in Sections II-A and II-B respectively. The choice of a speech model from ASR and its implications for the measure are discussed in Section II-C.

A. Text-Level Objective Intelligibility

A measure of objective intelligibility at the text level is established in this section. Let $\boldsymbol{F_x}$ and $\boldsymbol{F_n}$ represent some parametric descriptions of the speech signal and the environment noise respectively. We denote the estimate of $\boldsymbol{F_n}$ by $\mathcal{F}_{\boldsymbol{n}}$ and refer to it as the noise statistic. In addition, we denote the phonetic transcription of the message by $\boldsymbol{t}$, the rules for speech production by $\mathcal{P}$ and the rules for modifying speech production in the presence of noise by $\mathcal{R}$. The output of the speech production segment of the communication Markov chain can then be represented by the conditional probability density $\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}\right)$. The produced speech, thus, abides by a set of rules and adapts to the message and the environment.

The transmitted sound waveform reaches the listener contaminated by environment noise from the communication channel. Denoting some parametric description of the noisy signal by $\boldsymbol{F_y}$, we represent its distribution given the speech parameters and the noise statistic by $\mathrm{p}\left(\boldsymbol{F_y}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}}\right)$. The probability mass function of the decoded transcription $\boldsymbol{\tau}$, given $\boldsymbol{F_y}$ and a model $\mathcal{V}$ for decoding speech features, is $\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{F_y},\mathcal{V}\right)$. Assuming that $\boldsymbol{\tau}$ is conditionally independent

from $\boldsymbol{F_x}$ and $\mathcal{F}_{\boldsymbol{n}}$ given $\boldsymbol{F_y}$ we integrate out $\boldsymbol{F_y}$ from $\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{F_y},\mathcal{V}\right)$ as follows:

$$\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right)=\int_{\boldsymbol{F_y}}\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{F_y},\mathcal{V}\right)\mathrm{p}\left(\boldsymbol{F_y}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}}\right)d\boldsymbol{F_y}. \quad (1)$$

The conditional independence assumption is satisfied from an ASR perspective where only the noisy features are needed to obtain a transcription [26]. From the perspective of a human listener this assumption is an approximation due to the inherent capability of listeners to adapt to noisy environments. This form of adaptation is not considered in the present study.

If we, further, integrate out $\boldsymbol{F_x}$ from (1), we obtain:

$$\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{t},\mathcal{P},\mathcal{R},\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right)=\int_{\boldsymbol{F_y}}\int_{\boldsymbol{F_x}}\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{F_y},\mathcal{V}\right)$$
$$\mathrm{p}\left(\boldsymbol{F_y}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}}\right)\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}\right)d\boldsymbol{F_x}d\boldsymbol{F_y}. \quad (2)$$

To optimize intelligibility at the transcription level we first need to define a distortion measure $\mathrm{d}\left(\boldsymbol{\tau},\boldsymbol{t}\right)$ between the transcriptions of the transmitted and the decoded messages. Note that the same logic applies if intelligibility is optimized at another level of the communication chain. Next, we need to identify the mean distortion D by computing the expected value of $\mathrm{d}\left(\boldsymbol{\tau},\boldsymbol{t}\right)$ over the stochastic variables. Using that the transcription set is discrete following a finite alphabet size, the mean distortion is expressed as:

$$\mathrm{D}\left(\mathcal{P},\mathcal{R},\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right)=$$
$$=\sum_{\boldsymbol{t}}\sum_{\boldsymbol{\tau}}\mathrm{d}\left(\boldsymbol{\tau},\boldsymbol{t}\right)\mathrm{p}\left(\boldsymbol{\tau},\boldsymbol{t}|\mathcal{P},\mathcal{R},\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right) \quad (3)$$
$$=\sum_{\boldsymbol{t}}\sum_{\boldsymbol{\tau}}\mathrm{d}\left(\boldsymbol{\tau},\boldsymbol{t}\right)\mathrm{p}\left(\boldsymbol{\tau}|\boldsymbol{t},\mathcal{P},\mathcal{R},\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right)\mathrm{p}\left(\boldsymbol{t}\right). \quad (4)$$

To obtain (4) from (3) we assumed that $\boldsymbol{t}$ is conditionally independent of $\mathcal{F}_{\boldsymbol{n}}$. This is a weak assumption that is satisfied as long as the modification strategy does not include reformulation of the message.

Finally, we minimize (4) with respect to the speech adjustment rules $\mathcal{R}$:

$$\mathcal{R}=\underset{\mathcal{R}'}{\operatorname{argmin}}\ \mathrm{D}\left(\mathcal{P},\mathcal{R}',\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right). \quad (5)$$

Adopting a hit-or-miss distortion criterion we obtain a tangible problem formulation. The general equation (5) transforms into

$$\mathcal{R}=\underset{\mathcal{R}'}{\operatorname{argmax}}\sum_{\boldsymbol{t}}\mathrm{p}\left(\boldsymbol{\tau}=\boldsymbol{t}|\boldsymbol{t},\mathcal{P},\mathcal{R}',\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right)\mathrm{p}\left(\boldsymbol{t}\right). \quad (6)$$

Before continuing with the optimization of the mean distortion, let us first investigate the effect of the length of the utterance represented by transcription $\boldsymbol{t}$. As this length increases from a phoneme to a word, sentence, article and a book, the importance of the particular $\boldsymbol{t}$ decreases due to the ergodic [27] nature of the communication process. Thus, for a $\boldsymbol{t}$ with a large time span, averaging over the transcription space does not affect the optimization process significantly and can be omitted:

$$\mathcal{R}\approx\underset{\mathcal{R}'}{\operatorname{argmax}}\ \mathrm{p}\left(\boldsymbol{\tau}=\boldsymbol{t}|\boldsymbol{t},\mathcal{P},\mathcal{R}',\mathcal{F}_{\boldsymbol{n}},\mathcal{V}\right)$$
$$\approx\underset{\mathcal{R}'}{\operatorname{argmax}}\int_{\boldsymbol{F_y}}\int_{\boldsymbol{F_x}}\mathrm{p}\left(\boldsymbol{\tau}=\boldsymbol{t}|\boldsymbol{F_y},\mathcal{V}\right)\mathrm{p}\left(\boldsymbol{F_y}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}}\right)$$
$$\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}'\right)d\boldsymbol{F_x}d\boldsymbol{F_y}. \quad (7)$$

Decreasing the time span of the transcription $\boldsymbol{t}$, e.g., by performing modifications at the phoneme or the word level, undermines the ergodicity assumption. This impairs the accuracy of the approximation of (6) by (7) and leads to a distortion measure whose validity becomes more localized. The duration of the transcription $\boldsymbol{t}$ in (7), thus, presents a trade-off between the rate of the adaptation of the modification and the extent of the validity region of the optimal modification parameters.

Assuming further that the conditional distributions $\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}'\right)$ and $\mathrm{p}\left(\boldsymbol{F_y}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}}\right)$ are peaky, the optimization problem is accurately approximated by:

$$\mathcal{R}\approx\underset{\mathcal{R}'}{\operatorname{argmax}}\ \mathrm{p}\left(\boldsymbol{\tau}=\boldsymbol{t}|\widehat{\boldsymbol{F}}_{\boldsymbol{y}},\mathcal{V}\right), \quad (8)$$

where

$$\widehat{\boldsymbol{F}}_{\boldsymbol{y}}=\underset{\boldsymbol{F_y}}{\operatorname{argmax}}\ \mathrm{p}\left(\boldsymbol{F_y}|\widehat{\boldsymbol{F}}_{\boldsymbol{x}},\mathcal{F}_{\boldsymbol{n}}\right) \quad (9)$$

$$\widehat{\boldsymbol{F}}_{\boldsymbol{x}}=\underset{\boldsymbol{F_x}}{\operatorname{argmax}}\ \mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}'\right). \quad (10)$$

The assumption on the peaky character of $\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}'\right)$ is satisfied when a realization of the speech signal is available, which is the most common scenario. In this case the distribution becomes a Dirac delta impulse at $\widehat{\boldsymbol{F}}_{\boldsymbol{x}}$. In the context of TTS, if the speech waveform is obtained based on sampling speech parameters from their respective probability models, the variance of $\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}'\right)$ would be non-zero. Most commonly, however, parameter sampling is not employed in TTS due to speech quality considerations.

The distribution $\mathrm{p}\left(\boldsymbol{F_y}|\boldsymbol{F_x},\mathcal{F}_{\boldsymbol{n}}\right)$ will have a peaky character when $\mathrm{p}\left(\boldsymbol{F_x}|\boldsymbol{t},\mathcal{F}_{\boldsymbol{n}},\mathcal{P},\mathcal{R}'\right)$ is peaky and the mapping from the noise statistic and the clean signal features to the noisy signal features introduces little uncertainty. We note that by working with a noise statistic $\mathcal{F}_{\boldsymbol{n}}$ rather than the random variable $\boldsymbol{F_n}$ the uncertainty in the noise characteristic is not taken into account.

B. Discriminative Nature of the Objective Function

This section illustrates the discriminative nature of the proposed measure of objective intelligibility and provides an expression that facilitates further analysis. Maximizing the objective function from (8) minimizes the probability for classification error in the transcription space. The discriminative nature of this objective function is readily illustrated as follows. The probabilities of all possible transcriptions must add up to one:

$$\mathrm{p}\left(\boldsymbol{t}\mid\widehat{\boldsymbol{F}}_{\boldsymbol{y}},\mathcal{V}\right)+\sum_{\boldsymbol{\tau},\boldsymbol{\tau}\neq\boldsymbol{t}}\mathrm{p}\left(\boldsymbol{\tau}\mid\widehat{\boldsymbol{F}}_{\boldsymbol{y}},\mathcal{V}\right)=1. \quad (11)$$

Employing the above relation we observe that optimizing $\mathrm{p}\left(\boldsymbol{t}\mid\widehat{\boldsymbol{F}}_{\boldsymbol{y}},\mathcal{V}\right)$ is theoretically equivalent to optimizing

$$\mathcal{O}=\log\left(\frac{\mathrm{p}\left(\boldsymbol{t}\mid\widehat{\boldsymbol{F}}_{\boldsymbol{y}},\mathcal{V}\right)}{\sum_{\boldsymbol{\tau},\boldsymbol{\tau}\neq\boldsymbol{t}}\mathrm{p}\left(\boldsymbol{\tau}\mid\widehat{\boldsymbol{F}}_{\boldsymbol{y}},\mathcal{V}\right)}\right), \quad (12)$$

due to the monotonic relation between the argument of the log in (12) and $\mathrm{p}\left(\boldsymbol{t} \mid \widehat{\boldsymbol{F}}_{\boldsymbol{y}}, \mathcal{V}\right)$, and the monotonic nature of the log.

The formulation from (12) is convenient as it facilitates further simplification. Applying Bayes' rule to the probability of transcription $\boldsymbol{t}$ yields:

$$\mathrm{p}\left(\boldsymbol{t} \mid \widehat{\boldsymbol{F}}_{\boldsymbol{y}}, \mathcal{V}\right) = \frac{\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{t}, \mathcal{V}\right) \mathrm{p}\left(\boldsymbol{t} \mid \mathcal{V}\right)}{\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \mathcal{V}\right)}. \tag{13}$$

Substitution of the transcription probabilities in (12) using (13) leads to:

$$\begin{aligned}\mathcal{O} = &\log\left(\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{t}, \mathcal{V}\right)\right) + \log\left(\mathrm{p}\left(\boldsymbol{t} \mid \mathcal{V}\right)\right) \\ &-\log\left(\sum_{\boldsymbol{\tau}, \boldsymbol{\tau} \neq \boldsymbol{t}} \mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{\tau}, \mathcal{V}\right) \mathrm{p}\left(\boldsymbol{\tau} \mid \mathcal{V}\right)\right),\end{aligned} \tag{14}$$

The second term of (14) is not affected by the optimization and can be omitted. The third term can be approximated by assuming that $\mathrm{p}\left(\boldsymbol{\tau} \mid \mathcal{V}\right) = |\boldsymbol{\tau}|^{-1}, \forall \boldsymbol{\tau}$, where $|\boldsymbol{\tau}|$ is the cardinality of the set of all possible transcriptions. Making all transcriptions equally probable is a reasonable approximation for adverse noise conditions where little contextual information is available to the listener. Furthermore, it offers a computational advantage by allowing use of the forward algorithm [28] to evaluate the sum.

C. Speech Model Choice and Measure Implications

In this section we discuss the choice for the speech model $\mathcal{V}$ and its implications for the objective measure $\mathcal{O}$. Section II-B established that

$$\mathcal{O} \approx \log\left(\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{t}, \mathcal{V}\right)\right) - \log\left(\sum_{\boldsymbol{\tau}, \boldsymbol{\tau} \neq \boldsymbol{t}} \mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{\tau}, \mathcal{V}\right)\right). \tag{15}$$

To proceed with the evaluation of $\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{\tau}, \mathcal{V}\right), \forall \boldsymbol{\tau}$ the model $\mathcal{V}$ needs to be specified. In this work we take $\mathcal{V}$ to be the speech model from an ASR system pre-trained on clean speech. The motivation for this choice is to induce a modification behavior that results in clean modified speech that combined with the disturbance becomes similar to the reference clean speech in the parametric space where $\mathcal{V}$ was trained. Note that using the speech model from a missing data speech recognizer [18] will produce a conceptually different approach.

Speech models for ASR build upon the use of hidden Markov models (HMMs) [28]. Gaussian mixture models (GMMs) are employed to approximate the distributions of feature vectors associated with the states of the HMM. For large-vocabulary recognition systems the states represent phonemes or, in more sophisticated models, stages of a phoneme [26]. Transition probabilities between the states characterize phone durations. The feature vectors are extracted on a per-frame basis where the length of the frame and the degree of overlap between consecutive frames are system parameters.

We can now relate the parametric description $\widehat{\boldsymbol{F}}_{\boldsymbol{y}}$ of the noisy signal to the frame-based features by

$$\widehat{\boldsymbol{F}}_{\boldsymbol{y}} = \left[\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{1}, \cdots, \widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{J}\right], \tag{16}$$

where J is the number of frames in the time span of $\boldsymbol{t}$. For a particular state sequence $\boldsymbol{s}$, $\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}}, \boldsymbol{s} \mid \boldsymbol{\tau}, \mathcal{V}\right)$ is computed as:

$$\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}}, \boldsymbol{s} \mid \boldsymbol{\tau}, \mathcal{V}\right) = \prod_{j=1}^{J} \mathrm{p}\left(\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{j} \mid \boldsymbol{s}^{j}, \mathcal{V}\right) \mathrm{p}\left(\boldsymbol{s}^{j} \mid \boldsymbol{s}^{j-1}, \boldsymbol{\tau}, \mathcal{V}\right), \tag{17}$$

where j is the index over the frames in the modification window, $\boldsymbol{s}^{j}$ is the state associated with frame j, $\mathrm{p}\left(\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{j} \mid \boldsymbol{s}^{j}, \mathcal{V}\right)$ is the likelihood of features $\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{j}$ for state $\boldsymbol{s}^{j}$ and $\mathrm{p}\left(\boldsymbol{s}^{j} \mid \boldsymbol{s}^{j-1}, \boldsymbol{\tau}, \mathcal{V}\right)$ is the probability for transition from $\boldsymbol{s}^{j-1}$ to $\boldsymbol{s}^{j}$ given the transcription. The assumption of a first-order Markovian character of the speech signal in the state space of the recognition system, as indicated by the right-hand side of (17), results in a simple model and reduces the computational complexity for evaluating the probability of a state sequence. Complexity is an important design characteristic from the perspective of ASR and is particularly relevant in the context of the proposed speech modification algorithm. Marginalizing $\boldsymbol{s}$ from (17) results in $\mathrm{p}\left(\widehat{\boldsymbol{F}}_{\boldsymbol{y}} \mid \boldsymbol{\tau}, \mathcal{V}\right)$.

Note that, in (17), probability is normalized on a per-frame rather than a per-phoneme basis. The duration of individual phones within the time span of transcription $\boldsymbol{t}$ may, however, vary significantly, e.g., plosives and fricatives are generally shorter than vowels. From the perspective of speech pre-enhancement we need a mechanism that takes into account that the length of a phone is not proportional to its importance for intelligibility at the level of, e.g., the word or the sentence. We correct the measure resulting from the use of a speech model from ASR by normalizing the contribution of individual phonemes to a fixed phone duration.

Use of a speech model from ASR also has an implication on the complexity of the pre-enhancement method. State-of-the-art recognizers use context-dependent phoneme models. Consequently, the model number grows exponentially from, e.g., 39 (the number of phonemes in the English language) to, at most, 39^3. While the forward algorithm, e.g., [28], computes efficiently the second term of (15), the associated complexity is prohibitive for speech pre-enhancement in on-line applications.

We identify two distinctive strategies that, used separately or in combination, reduce the algorithm complexity. The first strategy is to simplify the phoneme models by dropping the context dependence. The number of phoneme models is then drastically reduced and the consecutive re-evaluation of the forward algorithm becomes feasible. The second strategy is to consider only a select subset of alternative transcriptions based, e.g., on prior knowledge of likely confusions. In the extreme case, all alternative transcriptions can be omitted leading to an absolute as opposed to a discriminative objective measure. The accuracy of this approximation increases with the separation between the correct and the alternative transcriptions in the model space. This is, e.g., the case when the noise level decreases.

In this work we adopt the second complexity reduction strategy and focus on the maximization of the absolute measure. The latter represents a natural starting point and a suitable benchmark for further sophistication. Applying the phone-duration invariance correction that we motivated previously, the optimization problem takes the form:

$$\underset{\mathcal{R}'}{\operatorname{argmax}} \sum_{l=1}^{L} \sum_{j=1}^{J_l} J_l^{-1} \log \left(\frac{\mathrm{p}\left(\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{j} \mid \boldsymbol{s}^{j}, \mathcal{V}\right)}{\mathrm{p}\left(\boldsymbol{s}^{j} \mid \boldsymbol{s}^{j-1}, \boldsymbol{t}^{l}, \mathcal{V}\right)^{-1}} \right), \quad (18)$$

where l is a phoneme index, J_l denotes the number of frames spanned by phone l, j becomes an index over the frames within a phone and $\boldsymbol{t}^l$ is phoneme l in transcription $\boldsymbol{t}$. A single state sequence is considered in (18) to acknowledge that, for the correct transcription, the association between states and frames is known *a priori*. For future reference we denote the objective function from (18) by $\hat{\mathcal{O}}$.

We note that, throughout the derivation of the objective measure, the nature of the distortion was never specified. The implication is that the measure is general and, in principle, can accommodate both additive and convolutive distortions.

III. System operation

The principle of operation of the speech pre-enhancement system is described in Section III-A. Two specific speech modification strategies are presented in Section III-B. The complexity analysis and memory requirements of the method are given in Section III-C. System implementation considerations are presented in Section III-D.

A. Principle of Operation

The principle of operation of the proposed speech pre-enhancement system is presented in this section. Figure 2 provides a high-level diagram of the system. The essence of the objective function $\hat{\mathcal{O}}$ is illustrated by the *Objective Intelligibility Measure* block. For simplicity, the phone models are represented as single-state. In practice they are composed of three distinctive states, where each state is represented by a GMM. The correction for phone duration invariance, cf. Section II-C, is not depicted. The parameters α, β and γ represent the number of frames for each of the three phones in the word that is modified.

The computation of the terms $\mathrm{p}\left(\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{j} \mid \boldsymbol{s}^{j}, \mathcal{V}\right)$ in (18) requires the alignment between states $\boldsymbol{s}^j$ and feature vectors $\widehat{\boldsymbol{f}}_{\boldsymbol{y}}^{j}$. In text-to-speech this information is *a priori* available. For recorded speech, it must be inferred given the signal $\boldsymbol{x}$, the transcription $\boldsymbol{t}$ and the model $\mathcal{V}$. Forced-alignment [26] provides this information in an automated way. Performed off-line, i.e., prior to the speech modification, it does not have an impact on the complexity of the enhancement algorithm. Forced-alignment is not an error-free process and segmentation errors occur. Factors such as speaking rate and speaker accent affect its performance. Multiple hypothesis (state sequences) are generated during forced alignment. Of these, we consider only the most likely (*Viterbi*) sequence.

The noise statistic $\mathcal{F}_{\boldsymbol{n}}$ and the point-estimate $\widehat{\boldsymbol{F}}_{\boldsymbol{x}}$ of the features of the modified speech are used to obtain the point estimate $\widehat{\boldsymbol{F}}_{\boldsymbol{y}}$ of the features of the noisy modified speech. As indicated by Equation (16), the noisy signal features are extracted on a per-frame basis. The feature set is the same as the one used for training the speech model in the ASR system. ASR-training related details are provided in Section IV-A.

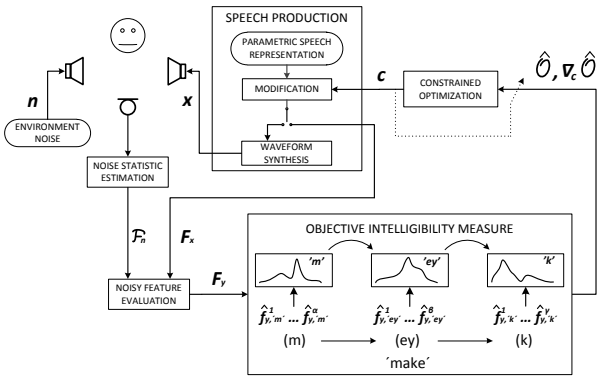

Fig. 2. Operation diagram of the proposed method.

Let us represent the effect of the speech adjustment rules $\mathcal{R}$ by the set of coefficients $\boldsymbol{c}$. The physical meaning of these coefficients depends on the modification strategy. The goal is to evaluate the objective function and its gradient with respect to $\boldsymbol{c}$. The optimizer adjusts the modification parameters and re-evaluates the objective function and the gradient. The process continues until a local optimum of the objective function is reached. An energy-preservation constraint, cf. Section III-B, applies during optimization. Upon convergence, the optimal adjustment rules are used to produce the speech presented to the listener.

B. Signal Modifications

The two modification strategies used for the validation of the proposed text-level objective intelligibility measure are described in this section. We argued, in Section II-A, that an increasing time span of transcription $\boldsymbol{t}$ gradually leads to message-independent modification parameters. Conversely, a decreasing time span of $\boldsymbol{t}$ reduces the general optimality of the modification parameters. A shorter duration of the modification window, however, entails a lower algorithmic delay and a faster adaptation to possibly changing noise properties. In this study we adopt a modification window of length equal to the word duration.

To facilitate comparison with the reference system, we consider two modifications - a) gain adjustment of band-energies in the channels of an auditory filter-bank and b) gain adjustment of phone-energies - and apply these to recorded speech. Note that neither of the two modifications affects the time scale of the speech signal. It follows that the denominator in the argument of the log in (18) has no effect on the optimization process and can be disregarded.

a) Gain adjustment of band energies: The band-energy modification controls the energy gain in the channels of a discrete Fourier transform (DFT) filter-bank over the duration

of a word. The segment of the signal waveform, which is aligned with the particular word is processed by a single DFT of adaptive length. The resulting spectrum is split into non-overlapping bands that are linearly-spaced on a Mel scale to reflect the decrease in frequency resolution with the increase in frequency in the human auditory system [29]. The effect of the rules $\mathcal{R}$ is represented by the set $\boldsymbol{c}_b^{\mathrm{T}} = [c_{b,1}, \cdots, c_{b,K}]$ of energy-scale coefficients, where the suffix b stands for *band-based* and K is the number of bands. Denoting the band energies in the original signal, for the duration of the modification window, by $\boldsymbol{e}_b^{\mathrm{T}} = [e_{b,1}, \cdots, e_{b,K}]$, an energy preservation constraint is defined as:

$$\boldsymbol{c}_b^{\mathrm{T}} \boldsymbol{e}_b = \boldsymbol{e}_b^{\mathrm{T}} \mathbf{1}, \;\; \boldsymbol{c}_b \geq 0, \tag{19}$$

where $\mathbf{1} \in \mathbb{R}^K$ is a vector of ones.

b) Gain adjustment of phone energies: The phone-energy modification adjusts the energy of each phone in a word under an energy preservation constraint. The effect of the rules $\mathcal{R}$ is, similarly, represented by the magnitudes of a set of energy-scale coefficients $\boldsymbol{c}_p^{\mathrm{T}} = [c_{p,1}, \cdots, c_{p,L}]$, where the suffix p represents *phone-based* and L is the number of phonemes in the word. Denoting the phone energies in the original signal by $\boldsymbol{e}_p^{\mathrm{T}} = [e_{p,1}, \cdots, e_{p,L}]$ an energy preservation constraint is defined as:

$$\boldsymbol{c}_p^{\mathrm{T}} \boldsymbol{e}_p = \boldsymbol{e}_p^{\mathrm{T}} \mathbf{1}, \;\; \boldsymbol{c}_p \geq 0. \tag{20}$$

Note that temporal gain modifications leading to large deviations in the energy dimension of the speech model feature space are not favored by the measure. Such modifications, however, can be beneficial to improving intelligibility. To reduce the sensitivity of the measure to such deviations we adjust the frame log-energy feature. Denoting the energy of speech frame j before modification by e^j, the energy of this frame after modification by $e^j_{\boldsymbol{x}}$ and the energy of the corresponding noisy speech frame by $e^j_{\boldsymbol{y}}$, the adjusted feature is obtained as

$$\log\left(e^j_{\boldsymbol{y}}\right)_a = \log\left(e^j_{\boldsymbol{y}}\right) - \left(\log\left(e^j_{\boldsymbol{x}}\right) - \log\left(e^j\right)\right). \tag{21}$$

Audible artifacts may occur in the modified speech due to the finite frequency and time resolution. Care must be taken to suppress these and avoid speech quality deterioration. To avoid tonal artifacts in the band-energy modification, we smooth the DFT-bin-level spectral gain by convolving it with a rectangular window. Discontinuities at the boundaries between phones, in the phone-energy modification, are de-emphasized by linear smoothing of overlapped boundary segments.

For theoretical optimality, multiple signal modifications should be optimized simultaneously. In practice, however, this is a challenging task for reasons that include i) the operation of the modifications on different time scales, ii) the increase in complexity resulting from the higher dimensionality of the optimization space and iii) convexity issues. We perform the band-energy and the phone-gain modifications in sequence, where the second modification acts on the output from the first.

C. Optimization Complexity and Memory Requirements

The algorithmic complexity of the proposed method and its memory requirements are analyzed in this section.

c) Computational complexity: On-line optimization of $\hat{\mathcal{O}}$, for the parameters $\boldsymbol{c}$ of the modifications discussed in Section III-B, requires the efficient evaluation of the objective function and its gradient. It is possible to derive an analytical and closed-form expression, which approximates closely the dependence of the gradient on $\boldsymbol{c}$ for the considered modifications. We adopt an alternative and generally-applicable approach, feasible in view of the relatively low dimensionality of the optimization problem, and approximate the gradient with finite differences [30].

The algorithmic complexity of the method, when applied in combination with the band-energy modification, scales as $O\left[\left(Z\log\left(Z\right) + 2\left|\widehat{\boldsymbol{f}}\right|\left|\boldsymbol{m}\right|\right) J\, I\, K\right]$, where Z is the number of bands in the analysis filter-bank of the recognizer front-end [26], $\left|\widehat{\boldsymbol{f}}\right|$ is the size of the feature vector, $|\boldsymbol{m}|$ is the number of components in a GMM and I is the number of iterations of the optimizer. The first term reflects the computation of the features in a frame given the power spectra of the speech and the noise. The second term represents the evaluation of (18) for one frame given the features. The common factor includes the number of frames J, the number of iterations I and the number of control parameters K to account for the finite-differences gradient approximation.

Similarly, the algorithmic complexity of the method with the phone-energy modification scales as $O\left[\left(Z\log\left(Z\right) + 2\left|\widehat{\boldsymbol{f}}\right|\left|\boldsymbol{m}\right|\right) J\, I\, L\right]$, where L is the number of phonemes in the modification window. On-line application of the method, with both modifications, is facilitated by adjusting the parameters I and K.

d) Memory requirements: The memory requirements of the algorithm can be split into a passive and an active share. The passive share is dominated by the acoustic speech models obtained from an ASR system. For context-dependent recognition systems the number of three-state HMMs is on the order of 16 K. Each state is represented by a GMM and a transition probability matrix. The state GMMs and transition matrices are pooled reducing greatly the memory footprint. Physically, 3.7 K GMMs need to be stored together with 41 transition matrices of dimension three. Using diagonal covariance matrices and $|\boldsymbol{m}|$ components per mixture requires the storage of $O\left[3700\;|\boldsymbol{m}|\;2\left|\widehat{\boldsymbol{f}}\right|\right]$ numbers, where $\left|\widehat{\boldsymbol{f}}\right|$ was defined as the dimension of the feature vector.

The active share of the memory footprint reflects the memory usage for the modification of a particular word. It includes the set of GMMs corresponding to the phonemes in this word. The spectrum of the word waveform and the modification parameters are stored as well.

D. System Implementation Considerations

This section presents the main design considerations needed to implement the proposed system. We consider modification-related settings, optimization initialization and noise estimation.

e) Modification-Related Settings : We use a filter-bank with 40 channels for the band-energy modification. This setting avoids distortion resulting from non-linear effects related to low spectral resolution. It also facilitates the comparison of results with the reference system, which uses a filter-bank with the same number of channels. The length of the rectangular window, used for smoothing the DFT-bin-level spectral gain, is $2\,\%$ of the DFT size.

The phone-gain modification is used in combination with the frame energy adjustment technique from (21). The size of the smoothing window applied to overlapping segments at the boundary of two phones is $15\,\mathrm{ms}$. This setting corresponds to the overlap between adjacent frames for feature extraction, cf. Section IV-A.

f) Optimization initialization: Two optimization problems are solved to obtain each modified word segment. We use a flat start for the phone-gain control parameters. The initialization for the band-gain control parameters is inspired by the shape of the hearing threshold [29]. The initial point is obtained by first concatenating two lines such that a peak is obtained at the band index associated with a frequency of $2300\,\mathrm{Hz}$. The increment between adjacent points is 1 for the line with positive slope and 0.3 for the line with negative slope. We scale the resulting series by 0.3 and compute their exponent. Finally, we normalize the initial point such that the energy preservation constraint is satisfied.

g) Noise estimation: The noisy features $\widehat{\boldsymbol{F}}_{\boldsymbol{y}}$ represent what the human listener ultimately observes. To optimize the modification parameters we need an estimate of the noisy features $\widehat{\boldsymbol{F}}_{\boldsymbol{y}}$ for each of a relevant set of modification parameter settings. For this purpose we create a set of noisy signals. We sum for each the realization of the modified clean signal and a realization of the noise signal that is equivalent to the noise signal expected in the considered scenario. In practice we use an excerpt of the noise signal that was played out $100\,\mathrm{ms}$ before the word. This effectively is an arbitrary realization of the noise signal generated based on its power spectrum and any other realization thus generated would give an equivalent result.

IV. Experimental Results

Training of the ASR system, providing the speech model, is discussed in Section IV-A. The reference system is presented in Section IV-B. The listening-test set-up and results are described in Sections IV-C and IV-D respectively, followed by an analysis of the method behavior in Section IV-E.

A. ASR Training

We trained a conventional HTK-based [26] speech recognizer. The training data consisted of 7138 utterances from the Wall Street Journal (WSJ0) database [31], at a sampling frequency of $16\,\mathrm{kHz}$. The analysis frame length was $25\,\mathrm{ms}$ and the update frame length was $10\,\mathrm{ms}$.

The feature set consisted of twelve MFCCs and the raw log-energy, together with the first and second-order differentials, i.e., a total of 39 features per frame. Cepstral mean normalization (CMN) [26] and energy normalization were applied over the duration of each utterance. The trained phoneme-models were context-dependent and consisted of three states each. A GMM with a diagonal covariance matrix and eight components was used to model each state.

We used the CMU dictionary (ver. 0.6) [32] to obtain phoneme-level transcriptions for the words in each utterance. The resulting ASR system achieved a word-correctness rate of $93.82\,\%$ for the read speech from the 5000-word closed-vocabulary task with non-verbalized punctuation [33] in the absence of noise.

B. Reference System

The reference system [3] was chosen based on its state-of-art performance and comparable speech modification capabilities. [3] performs spectro-temporal energy reallocation in speech-active regions based on the optimization of a perceptual distortion measure. Band energies are adjusted at the frame-level over the span of a sequence of frames, i.e., the energy gain is frame and band specific. The noise power spectrum is estimated using [34]. The method is transcription-independent and provides a closed-form solution for the optimal modification parameters under an energy-preservation constraint.

The following settings were used for the reference method. The auditory filter-bank contains 40 channels with the cut-off frequency of the highest channel at $8\,\mathrm{kHz}$. The size of a frame is $32\,\mathrm{ms}$ and the overlap between adjacent frames is $50\,\%$. To facilitate comparison of the results with the proposed method we set the algorithmic delay to $350\,\mathrm{ms}$, reflecting the average word duration in the test material. Energy reallocation was, thus, performed over the span of 20 frames.

C. Listening Test Set-up

This section describes the set-up for the listening test. The test conditions included speech-shaped and multi-speaker babble noise [35] at SNR levels of $-4\,\mathrm{dB}$ and $-9\,\mathrm{dB}$. The speech material was composed of 12 sets (50 to 61) of ten sentences from the Harvard sentence database [36]. The recordings come from the ITU-T database [37] and are pronounced by a male, native American-English speaker. This choice of material is motivated with the broad use of [36] for intelligibility and quality assessment studies and the availability of [37]. The beginning and the end of each utterance were padded with $500\,\mathrm{ms}$ of silence. The noise was present for the full duration of the signal, but the SNR was established based on the effective speech duration. The noise signals were extracted from long noise recordings using a random starting point for each speech signal.

Twelve native English speakers, aged 18 to 38, were recruited and compensated financially for their participation. Each participant adjusted the listening level, before the test, based on demo material presented in speech-shaped and babble noise at $-4\,\mathrm{dB}\,\mathrm{SNR}$. Test participation required 35 minutes on average.

The test was conducted in silent conditions using a computer-based application with a rudimentary text interface and a pair of Beyerdynamic DT 770 headphones. Each subject listened to a single presentation of each of the 120 sentences.

The playback was initiated by the listener. After each presentation, the test participant was prompted to type in the perceived utterance using special characters in place of unidentified words.

The speech material was evenly distributed among the four test conditions. As a result, each condition was represented by three sets of ten sentences. Within a condition, one set of sentences represented the natural speech, another set represented speech modified by the reference method and the remaining set represented speech modified by the proposed method. The presentation order for the proposed and the reference methods was reversed in half of the test material to reduce presentation order effects. Each set of sentences was presented with each combination of processing method and condition over the entire listening test.

D. Subjective Evaluation Results

Subjective recognition rates were computed from the listener's input for each utterance as the ratio of the correctly identified to the total number of words (including prepositions and articles). The per-set recognition rates, computed by averaging the per-sentence recognition rates over a sentence set and listener, for the speech-shaped and the babble noise conditions are presented in Table I and Table II respectively. We use the abbreviation Nat. for natural speech, PD for the reference method and Prop. for the proposed method. The lowest recognition rates, in each condition, always occurred for the natural speech. A per-set recognition rate of zero was observed on one occasion only for natural speech.

TABLE I
SUBJECTIVE (PER-SET) RECOGNITION RATES FOR SPEECH-SHAPED NOISE.

SNR	-4 dB			-9 dB		
Subject	**Nat.**	**PD**	**Prop.**	**Nat.**	**PD**	**Prop.**
1	0.63	0.93	0.91	0.17	0.28	0.68
2	0.67	0.87	0.91	0.15	0.62	0.43
3	0.51	0.55	0.73	0.11	0.14	0.46
4	0.82	0.78	0.88	0.12	0.57	0.76
5	0.39	0.9	0.83	0.19	0.25	0.38
6	0.73	0.85	0.94	0.08	0.43	0.64
7	0.74	0.95	0.9	0.18	0.53	0.68
8	0.61	0.81	0.85	0.12	0.25	0.43
9	0.49	0.85	0.72	0.16	0.34	0.5
10	0.7	0.94	0.89	0.27	0.64	0.86
11	0.75	0.98	0.94	0.32	0.34	0.7
12	0.58	0.9	0.89	0.16	0.57	0.78
mean	**0.64**	**0.86**	**0.87**	**0.17**	**0.41**	**0.61**
std	**0.12**	**0.11**	**0.07**	**0.06**	**0.16**	**0.15**

The mean and the standard deviation of the per-set recognition rates over each approach and condition are presented in bold font at the bottom of Table I and Table II. The per-condition results are also visualized in Figure 3. On average, the proposed algorithm outperforms the reference system, which in turn produces speech that is more intelligible than natural speech in noise.

Note that the performance with either of the three approaches is lower for babble noise than for speech-shaped noise. This behavior can be explained in terms of the informational masking effect [38]. In particular, the babble noise recording contains dominant voices from speakers located closer to the recording device. These voices have a distracting effect on the listener during a test.

TABLE II
SUBJECTIVE (PER-SET) RECOGNITION RATES FOR BABBLE NOISE.

SNR	-4 dB			-9 dB		
Subject	**Nat.**	**PD**	**Prop.**	**Nat.**	**PD**	**Prop.**
1	0.37	0.71	0.88	0.05	0.35	0.44
2	0.56	0.83	0.81	0.09	0.29	0.53
3	0.31	0.44	0.68	0.09	0.2	0.21
4	0.56	0.74	0.82	0.07	0.24	0.51
5	0.4	0.79	0.8	0.08	0.21	0.25
6	0.38	0.91	0.82	0.04	0.3	0.4
7	0.52	0.81	0.82	0.09	0.16	0.48
8	0.36	0.91	0.78	0.01	0.1	0.36
9	0.34	0.55	0.66	0.09	0.07	0.16
10	0.61	0.88	0.87	0.01	0.36	0.54
11	0.48	0.91	0.95	0	0.21	0.49
12	0.33	0.79	0.82	0.11	0.27	0.5
mean	**0.44**	**0.77**	**0.81**	**0.06**	**0.23**	**0.41**
std	**0.1**	**0.14**	**0.08**	**0.03**	**0.09**	**0.13**

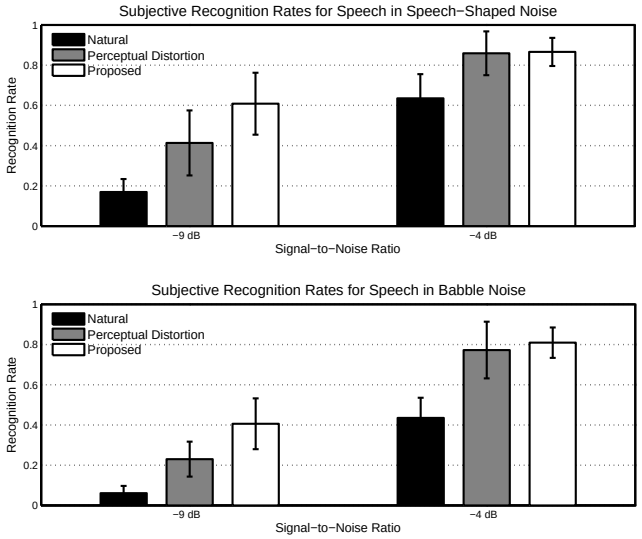

Fig. 3. Subjective intelligibility of natural and pre-processed speech in speech-shaped (top) and babble (bottom) noise.

The statistical significance of the intelligibility improvement from speech modification was evaluated using the Wilcoxon signed rank test [39]. This choice of test avoids assumptions on the distribution of the recognitions rates. The results are presented in Table III, where BBL denotes babble and SSN denotes speech-shaped noise. In all conditions, both modification methods produce significant intelligibility improvement over natural speech. The proposed algorithm is significantly better than the reference system for both types of noise at $-9\,\mathrm{dB\,SNR}$. At $-4\,\mathrm{dB\,SNR}$ the significance of the improvement over the reference system decreases. This is the expected behavior as the average recognition rates approach the performance limit.

TABLE III
WILCOXON SIGNED RANK TEST SIGNIFICANCE ANALYSIS.

SSN, -4 dB	Nat.	PD	Prop.	SSN, -9 dB	Nat.	PD	Prop.
Nat.	1	0.001	0.001	Nat.	1	0.001	0.001
PD	-	1	1	PD	-	1	0.005
Prop.	-	-	1	Prop.	-	-	1
BBL, -4 dB	Nat.	PD	Prop.	BBL, -9 dB	Nat.	PD	Prop.
Nat.	1	0.001	0.001	Nat.	1	0.001	0.001
PD	-	1	0.275	PD	-	1	0.001
Prop.	-	-	1	Prop.	-	-	1

E. Method Behavior

The effect of the proposed algorithm on the output speech waveform is illustrated for sentence three from set 50 and sentence one from set 51 in [36]. The first sentence is modified for presentation in speech-shaped noise at $-4\,\mathrm{dB}\,\mathrm{SNR}$, while the second sentence is modified for presentation in babble noise at $-4\,\mathrm{dB}\,\mathrm{SNR}$.

Figure 4 shows the waveform of the first sentence before modification (top), and after each of the two modifications (middle and bottom). To facilitate the interpretation, the phonetic transcription of the sentence is inserted in the top plot. The phonemes are aligned with the position of the corresponding phones within the utterance. The original sentence is *"Their eyelids droop for want of sleep."*. The modification parameters for the band-energy modification are visualized by means of time-frequency tiles in the top plot of Figure 5. Full-band tiles are used to illustrate the phone-gain modification in the bottom plot of the same figure. In both cases, darker color indicates higher gain. The phone-gains are visualized after a log transform to reduce loss of visual information due to large gain differences. The spectrograms of the clean and the noisy natural and modified (after both modifications) speech are presented in Figure 6.

The signal waveform after the band-energy modification suggests temporal energy distribution within the modification window, e.g., the phone *"f"* in *"f-er"* and the phone *"t"* in *"w-aa-n-t"* increase in amplitude. The observed behavior is the result of an energy transfer to a frequency range dominated by particular phones. After re-synthesis, such phones appear amplified while others indicate the opposite effect. We recall that the band-energy modification operates simultaneously over the sequence of phonemes forming the word.

The effect of the phone-gain modification is clearly visible for, among others, the phone *"z"* in *"ay-l-ih-d-z"* and the phone *"p"* in *"s-l-iy-p"*. The outcome from the phone-gain modification is suggestive of an energy transfer from vowels to consonants. A dependence of the behavior of this modification on the duration of the phones is not observed. We recall that avoiding such dependence was the motivation for performing phone-duration normalization.

Using the same approach, the behavior of the modifications is illustrated for the sentence *"Shake the dust from your shoes stranger."* in Figures 7,8 and 9. The optimization in this case is performed for presentation in multi-speaker babble noise. The observed trends are similar to those in the previous example. Both the band-gain and the phone-gain modifications appear to strengthen the weaker phonemes at the expense of the stronger ones. This behavior often leads to an energy transfer from vowels to consonants.

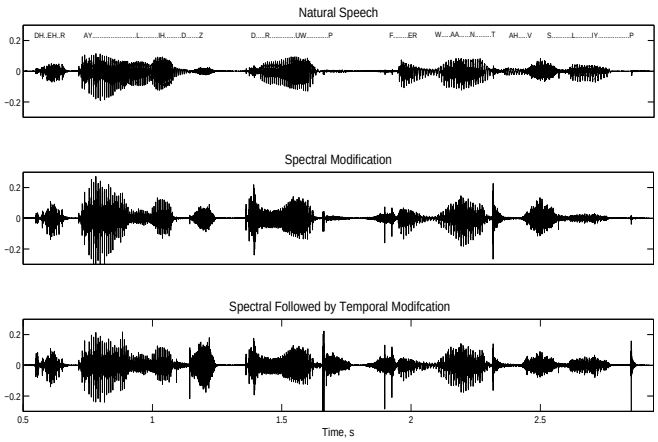

Fig. 4. Effect of the speech modifications on the signal waveform for speech-shaped noise at -4dB.

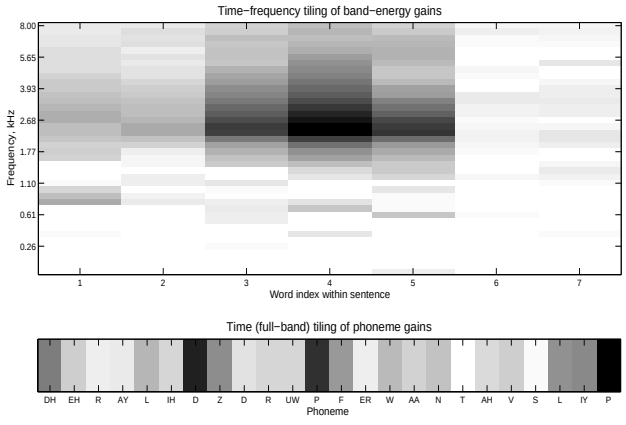

Fig. 5. Method behaviour for the sentence "Their eyelids droop for want of sleep.". The optimization is performed for presentation in $-4\,\mathrm{dB}\,\mathrm{SNR}$ speech-shaped noise.

V. CONCLUSIONS

A practical and effective approach to speech pre-enhancement for improving intelligibility in noisy environments can be established by optimizing objective intelligibility at the level of a phonetic transcription. The proposed framework is expected to work well for a range of speech modification strategies providing versatility that can be exploited in various application scenarios. The extent of this range is a topic for future studies. We substituted the theoretically-optimal discriminative measure with a non-discriminative measure due to complexity considerations. This approximation suggests the potential for further improvement of the method.

ACKNOWLEDGMENT

The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme

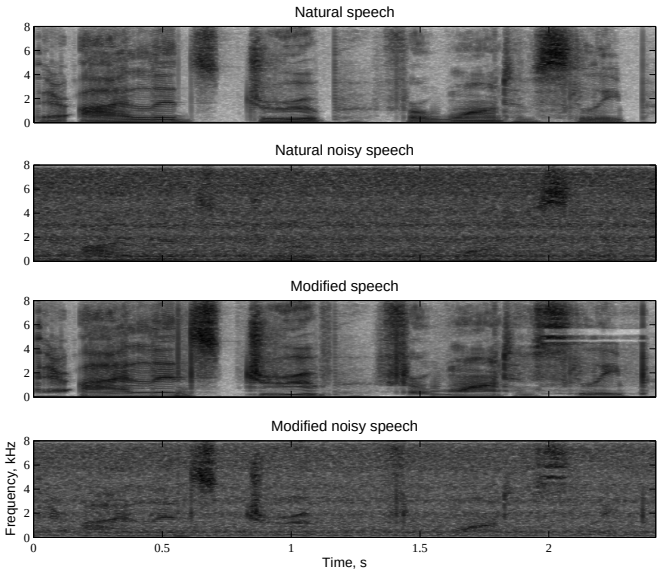

Fig. 6. Spectrograms for the natural and modified signals of the sentence "Their eyelids droop for want of sleep." with and without speech-shaped noise at $-4\,\mathrm{dB\,SNR}$. The beginning and trailing silence periods have been omitted.

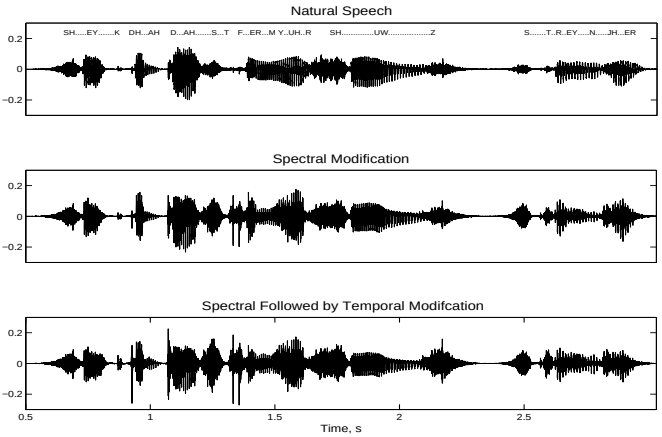

Fig. 7. Effect of the speech modifications on the signal waveform for multi-speaker babble noise at $-4\mathrm{dB\,SNR}$.

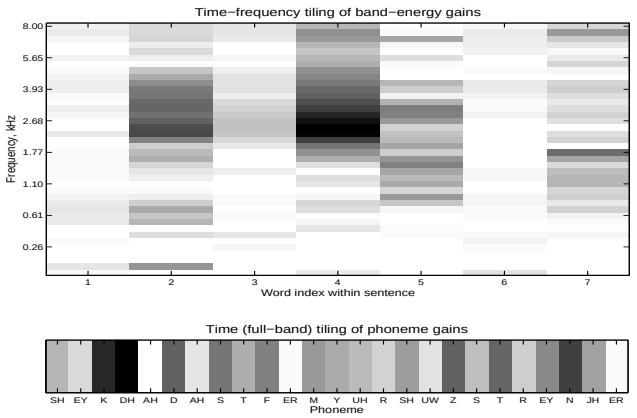

Fig. 8. Method behaviour for the sentence "Shake the dust from your shoes stranger.". The optimization is performed for presentation in $-4\,\mathrm{dB\,SNR}$ multi-speaker babble noise.

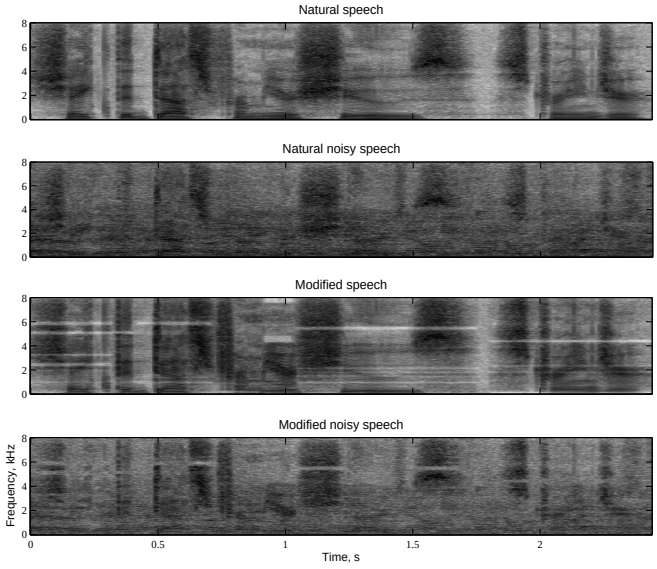

Fig. 9. Spectrograms for the natural and modified signals of the sentence "Shake the dust from your shoes stranger." with and without multi-speaker babble noise at $-4\,\mathrm{dB\,SNR}$. The beginning and trailing silence periods have been omitted.

within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230.

REFERENCES

[1] V. Hazan and A. Simpson, "The Effect of Cue-Enhancement on Consonant Intelligibility in Noise: Speaker and Listener Effects," *Language and Speech*, vol. 43, pp. 273–294, 2000.

[2] B. Sauert, G. Enzner, and P. Vary, "Near End Listening Enhancement with Strict Loudspeaker Output Power Constraint," in *Intern. Workshop on Acoustic Echo and Noise Control*, 2006.

[3] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A Speech Preprocessing Strategy for Intelligibility Improvement in Noise Based on a Perceptual Distortion Measure," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2012, pp. 4061–4064.

[4] N. Hodoshima, T. Arai, A. Kusumoto, and K. Kinoshita, "Improving Syllable Identification by a Preprocessing Method Reducing Overlap-Masking in Reverberant Environments." *J. Acoust. Soc. Am.*, vol. 119, pp. 4055–4064, 2006.

[5] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of Noise on Speech Production: Acoustic and Perceptual Analyses." *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, Sep 1988.

[6] J. C. Junqua, "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

[7] Y. Lu and M. Cooke, "Speech Production Modifications Produced by Competing Talkers, Babble, and Stationary Noise." *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, Nov 2008. [Online]. Available: http://dx.doi.org/10.1121/1.2990705

[8] B. Eukel, "Phonotactic Basis for Word Frequency Effects: Implications for Lexical Distance Metrics," *J. Acoust. Soc. Am.*, vol. 68, p. s33, 1980.

[9] R. Patel and K. W. Schell, "The Influence of Linguistic Content on the Lombard Effect," *J. Speech, Lang., and Hear. Res.*, vol. 51, pp. 209–220, 2008.

[10] S. Yoo, J. R. Boston, J. D. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C. C. Li, "Speech Enhancement Based on Transient

Speech Information," in *Proc. Appl. Sig. Proc. Audio and Acoust. Workshop*, 2005, pp. 62–65.
[11] P. S. Chanda and S. Park, "Speech Inteligibility Enhancement Using Tunable Equalization Filter," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2007, pp. 613–616.
[12] J. W. Shin, W. Lim, J. Sung, and N. S. Kim, "Speech Reinforcement Based on Partial Specific Loudness," in *Proc. Interspeech*, 2007, pp. 978–981.
[13] B. Sauert, H. W. Löllmann, and P. Vary, "Near End Listening Enhancement by Means of Warped Low Delay Filter-Banks," in *Proc. Voice Communication Conf.*, 2008, pp. 1–4.
[14] B. Sauert and P. Vary, "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index and Audio Power Limitations," in *Proc. Europ. Sig. Proc. Conf.*, 2010, pp. 1919–1923.
[15] Y. Tang and M. Cooke, "Energy Reallocation Strategies for Speech Enhancement in Known Noise Conditions," in *Proc. Interspeech*, 2010, pp. 1636–1639.
[16] C. V. Botinhao, R. Maia, J. Yamagishi, S. King, and H. Zen, "Cepstral Analysis Based on the Glimpse Proportion Measure for Improving the Intelligibility of HMM-based Synthetic Speech in Noise," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2012.
[17] American National Standard, "Methods for the Calculation of the Speech Intelligibility Index," 1997.
[18] M. Cooke, "A Glimpsing Model of Speech Perception in Noise." *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, Mar 2006.
[19] H. Fletcher, *Speech and Hearing in Communication*, J. B. Allen, Ed. Acoust. Soc. Am., 1995.
[20] J. B. Allen, "How do Humans Process and Recognize Speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 567–577, 1994.
[21] U. G. Goldstein, "An Articulatory Model for the Vocal Tracts of Growing Children." Ph.D. dissertation, Massachusetts Institute of Technology, 1980.
[22] T. Crowley, *An Introduction to Historical Linguistics*. Oxford University Press, 1997.
[23] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech, and Lang. Proc.*, no. 99, 2011, early Access.
[24] P. N. Petkov, W. B. Kleijn, and G. E. Henter, "Enhancing Subjective Speech Intelligibility Using a Statistical Model of Speech," in *Proc. Interspeech*, 2012.
[25] ——, "Speech Intelligibility Enhancement Using a Statistical Model of Clean Speech," in *The Listening Talker*, 2012, p. 77.
[26] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.
[27] M. Brin and G. Stuck, *Introduction to Dynamical Systems*. Cambridge University Press, 2002.
[28] L. W. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[29] B. C. Moore, *An Introduction to the Psychology of Hearing*. Elsevier Academic Press, 2004.
[30] R. J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics (SIAM), 2007.
[31] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-Based CSR Corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.
[32] "The CMU Pronouncing Dictionary," Carnegie Mellon University, ftp://ftp.cs.cmu.edu/project/speech/dict/.
[33] H. Murveit, J. Butzberger, and M. Weintraub, "Performance of SRI's DECIPHER Speech Recognition System on DARPA's CSR Task," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 410–414.
[34] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based Noise PSD Tracking with Low Complexity," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2010, pp. 4266–4269.
[35] A. P. Varga, J. M. Steenneken, M. Tomlinson, and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," DRA Speech Research Unit, Tech. Rep., 1992.
[36] "IEEE Recommnded Practice for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.
[37] ITU-T Rec. P.Sup23, "ITU-T Coded-Speech database," 1998.
[38] N. I. Durlach, C. R. Mason, G. K. Jr., T. L. Arbogast, H. S. Colburn, and B. G. Shinn-Cunningham, "Note on Informational Masking (l)," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 2984–2987, 2003.
[39] D. F. Bauer, "Constructing Confidence Sets Using Rank Statistics," *J. Am. Stat. Assoc.*, vol. 67, pp. 687–690, 1972.

Petko N. Petkov received the B.Sc. degree in communication engineering from the Technical University of Sofia, Bulgaria and the M.Sc. degree in electrical engineering from KTH - The Royal Institute of Technology, Stockholm, Sweden. He is currently pursuing the Ph.D. degree at the Sound and Image Processing Lab, School of Electrical Engineering, KTH. He was a research and development engineer with Global IP Solutions in 2006-2007. Petko has been a visiting researcher at the Tampere University of Technology, Finland and at KTH. His research interests include the application of signal processing and machine learning to problems in speech and audio processing.

Gustav Eje Henter Gustav Eje Henter received his MSc in Engineering Physics in 2007 from KTH - The Royal Institute of Technology in Stockholm, Sweden. He is currently pursuing PhD studies at the Sound and Image Processing laboratory within the School of Electrical Engineering at KTH. In spring 2011 he was a visiting PhD student at the Communications and Signal Processing (CaSP) group at Victoria University of Wellington (VUW), New Zealand.

His current research interests include statistical models, especially nonparametric methods, and machine learning, particularly with applications to speech synthesis and related topics.

W. Bastiaan Kleijn Bastiaan Kleijn has been a Professor at Victoria University of Wellington since 2010. He is also a Professor at Delft University of Technology and at KTH in Stockholm, where he was the Head of the Sound and Image Processing Laboratory until he moved to New Zealand. Before joining KTH in 1996, he worked at AT&T Bell Laboratories (Research) on speech processing. He was a founder of Global IP Solutions, which developed voice and video processing engines for, among others, Google, Skype, and Yahoo and was sold to Google in 2010. Kleijn holds a Ph.D. in Electrical Engineering from Delft University of Technology and an MSEE from Stanford. He also earned a Ph.D. in Soil Science and an MS In Physic from the University of California, Riverside. He is a Fellow of the IEEE.