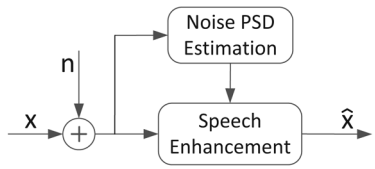
Spectral Dynamics Recovery for Enhanced Speech Intelligibility in Noise

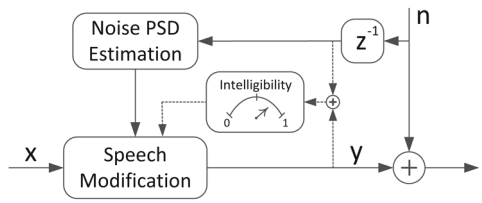
Petko N. Petkov, *Member, IEEE*, and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—Speech intelligibility in noisy environments decreases with an increase in the noise power. We hypothesize that the differences of subsequent short-term spectra of speech, which we collectively refer to as the speech *spectral dynamics*, can be used to characterize speech intelligibility. We propose a distortion measure to characterize the deviation of the dynamics of the noisy modified speech from the dynamics of natural speech. Optimizing this distortion measure, we derive a parametric relationship between the signal band-power before and after modification. The parametric nature of the solution ensures adaptation to the noise level, the speech statistics and a penalty on the power gain. A multi-band speech modification system based on the single-band optimal solution is designed under a total signal power constraint and evaluated in selected noise conditions. The results indicate that the proposed approach compares favorably to a reference method based on optimizing a measure of the speech intelligibility index. Very low computational complexity and high intelligibility gain make this an attractive approach for speech modification in a wide range of application scenarios.**

***Index Terms*—Environment adaptation, speech intelligibility enhancement, speech modification.**



Speech enhancement.



Intelligibility-enhancing speech modification.

Fig. 1. Problem formulation differences between speech enhancement and intelligibility-enhancing speech modification.

I. Introduction

THE degradation of speech intelligibility in noise is an extensively studied phenomenon, e.g., [1]–[5]. From an engineering perspective two problems related to the loss of intelligibility in noise exist and are actively worked on. On the one hand is the speech enhancement problem where it is assumed that the speech and the noise signals are only observed as a mixture, and the objective is to suppress the noise and improve the quality of service, e.g., [6]–[8]. The dual problem assumes that the information-bearing signal is known and the objective is to modify this signal such that its intelligibility, in the presence of noise, is enhanced. Fig. 1 illustrates the two scenarios. In this paper we focus on the problem of intelligibility-enhancing speech modification.

Manuscript received February 15, 2014; revised July 12, 2014; accepted December 04, 2014. Date of publication December 18, 2014; date of current version January 15, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rongshan Yu.

P. N. Petkov was with the KTH Royal Institute of Technology, 100 44 Stockholm, Sweden. He is now with the Cambridge Research Laboratory, Toshiba Research Europe Limited, Cambridge, CB4 0GZ, U.K. (e-mail: petko.petkov@crl.toshiba.co.uk).

W. B. Kleijn, was with the KTH Royal Institute of Technology, 100 44 Stockholm, Sweden. He is now with the Department of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand, and also with the Delft University of Technology, 2628 CN Delft, The Netherlands (e-mail: bastiaan.kleijn@ecs.vuw.ac.nz).

Digital Object Identifier 10.1109/TASLP.2014.2384271

Speech modification for improved intelligibility in noise is a problem with relatively long history. Recently, it has attracted increased interest not least due to the efforts within the Listening Talker (LISTA) project [9]. Methods for speech modification can be divided into two large classes: measure-driven and rule-driven.

Rule-based methods adopt a single or a combination of strategies known to improve speech intelligibility in noise. These methods do not typically adapt to changes in the environment. Absence of adaptation makes them highly robust and computationally efficient. Lack of adaptation, however, suggests that the extent of modification at any instant is either excessive or insufficient. A particularly successful line of work uses a combination of dynamic range compression (DRC) [10] and high-pass filtering [11]. This track has been exploited and augmented further in, among others, [12]where DRC is applied in combination with spectral shaping (SS). Rule-based methods using other signal characteristics and processing techniques include, e.g., [13], [14].

Measure-driven speech modification forms a recent addition to the range of methods addressing the deterioration of speech intelligibility in noise. The fundamental difference from rule-based methods is that the degree of modification (and possibly also the choice of modification strategy) is the outcome of optimizing a distortion measure. The resulting framework is adaptive and introduces feed-back on the expected intelligibility gain for any degree of modification. An attractive feature of measure-driven modification methods is that multiple modification strategies can be exploited simultaneously. The intelligibility

gain achievable with such methods is limited by the accuracy of the intelligibility measure and the choice of modification strategies. Additional challenges may be posed by convexity and computational complexity issues.

Within the class of measure-based methods, further sub-classification is possible based on a hierarchical outlook of the communication process [2]. Most methods use measures that operate at the power spectral level, which is a relatively low level in the communication hierarchy. Among these, [15], [16] optimize a measure based on the speech intelligibility index (SII) [3]. These adjust the signal-to-noise ratios (SNRs) in an auditory filter-bank using perceptual considerations represented by band-importance weights. An approach exploring both frequency-dependent and frequency-independent SNR recovery followed by output power limitation is discussed in [17]. Optimization of a perceptual distortion measure to derive band-specific power gains is pursued in [18]. [19], [20] optimize distortion measures derived from a glimpsing model of speech intelligibility [21].

An operating level at which phonetic information becomes visible is considered a high level. Using a high-level measure based on phoneme recognition accuracy offers the advantage of multi-modal modification and optimality at a level that more accurately reflects subjective intelligibility [22]. Higher computational complexity and possible robustness issues due to the dependence on additional information narrow the practical applicability of such methods.

To preserve, in part, the advantages of high-level measure optimization, and simultaneously increase robustness and computational efficiency, the projection of a high-level measure to the feature space is considered. Our objective is to improve the similarity between the noisy modified and the natural speech features from the feature set of a speech recognition system. Typically, Mel cepstra (static features) and their differences (dynamic features) [23] are used. Recent findings in neuroscience, which highlight the importance of the temporal evolution of the speech signal to intelligibility [24], [25], motivate us to focus on the dynamic features. As the cepstra are a unitary transform of smoothed log spectrum, this is equivalent to considering the dynamics of smoothed spectra.

In [26] we pursued the track of preserving the spectral dynamics down to a band-power threshold level. The method achieved an intelligibility gain over natural speech, but left room for improvement as a result of the inefficient use of signal power to preserve dynamics at high power levels with little positive impact on intelligibility. This issue is now addressed by introducing a distortion measure that characterizes the deviation of the instantaneous power dynamics of the noisy speech from the dynamics of the clean speech.

We propose an optimization criterion as the combination of the dynamics-distortion measure and a term that penalizes the band-power gain. Working with instantaneous power avoids complications from considering frame sequences and gives a parametric closed-form solution that relates the signal power before and after modification. The optimal power relations have a compressive characteristic. Autonomous adaptation of the power mapping functions to the statistics of speech and the environment noise generalizes the use of DRC for speech intelligibility enhancement.

Speech modification for improved intelligibility in noise is commonly performed under a power constraint to prevent damage to both the auditory system and the audio equipment. A power limit suggests that it is attractive to split the speech signal into bands and explore dynamics recovery trade-offs. In addition, multi-band processing allows better adaptation to the environment and modeling of the non-uniform spectral resolution of the auditory periphery [27].

The optimal relationship between the input and the output signal power is used to design a multi-band speech modification system. The optimization of the power distribution among individual bands is governed by the objective of global recovery of the spectral dynamics. In the following this approach is referred to as spectral dynamics recovery (SDR).

A tractable mathematical formulation is obtained by working with the marginal distributions of the (scalar) speech band-powers. Taking away the regularizing effect of the joint distribution, however, leads to a perceptual effect that limits the intelligibility gain for naive listeners. We address this issue by performing a perceptually-motivated adjustment to the modified band-powers. Experimental validation shows that the proposed method compares favorably to the SII-optimal speech modification approach from [16], under conditions commonly encountered in on-line communication scenarios. Higher, on average, intelligibility gain is achieved at a fraction of the computational complexity of the reference method and a shorter algorithmic delay.

The remainder of this paper is organized as follows. The theory behind the proposed approach is presented in Section II. Aspects related to the design of a practical multi-band speech modification system are discussed in Section III. Results from the objective and subjective evaluations of the method are given in Section IV followed by conclusions in Section V.

II. Theoretical Foundations

Recent studies in neurosciences emphasize the importance of the temporal structure of speech to the ability of humans to achieve high recognition even in severe noise conditions [24], [25], [28], [29]. Neuronal oscillations engage in tracking the speech dynamics on a multi-time-scale level. Successful tracking results in high speech intelligibility. We exploit the relevance of the dynamic properties of speech in developing a method for speech modification to improve its intelligibility in noise. Taking a simplified perspective, the focus is placed on increasing the similarity between the instantaneous power dynamics between clean and modified noisy speech. A distortion measure is first formulated to characterize the deviation. A criterion based on that measure is then optimized to derive a parametric relationship between the power before and after modification. Finally, a multi-band scenario is considered in which the free band-specific parameters are optimized for global dynamics recovery under a power constraint.

A. Single-Band Dynamics Recovery

A distortion measure that characterizes the deviation of the noisy power dynamics from the dynamics of clean speech is formulated in this section. An optimization criterion that combines the distortion measure and a term that penalizes the signal

power gain is proposed, and the general solution to the resulting optimization problem is presented. Choosing appropriate initial and final conditions, a particular parametric solution is derived and constrained to ensure monotonic behavior.

1) Problem Formulation: Let x denote a realization of the instantaneous band-power, i.e., the power in one band of the original speech signal, where for compactness of notation the band index is omitted. The modified band-power is denoted by y, where $y = y(x, n)$ and n is the power of the noise in that band. Taking into consideration the log-power sensitivity of the auditory periphery, the dynamics are recovered completely if the signal modification enforces the relation:

$$\frac{d\log(y+n)}{d\log(x)} = 1. \tag{1}$$

Complete dynamics recovery is neither power-efficient nor necessary to achieve good intelligibility. The focus is, therefore, placed on optimizing a distortion measure that characterizes the deviation of the dynamics of the modified noisy speech from the dynamics of the clean speech. The distortion is quantified as the squared distance of the left and right-hand sides of (1):

$$D_0 = \left(\frac{xy'}{y+n} - 1\right)^2, \tag{2}$$

where $y' = dy/dx$, to penalize large deviations of the dynamics as these have an adverse effect on intelligibility. To facilitate regulation of the consumed power, an optimization criterion that includes both (2) and a penalty on the power gain is formulated as:

$$\eta_0 = \int_\alpha^\beta \left(D_0 + \lambda n \frac{y}{x}\right) f_X(x)\, dx, \tag{3}$$

where α and β determine the power range in which optimal dynamics recovery is pursued, λ is a Lagrange multiplier and $f_X(x), x \in [\alpha, \infty)$ is the probability density function (PDF) of the instantaneous signal band-power. The lower bound α facilitates use of voice activity detection. The penalty term includes the factor n to ensure that in the absence of noise the signal is not modified. It is assumed that the noise power spectral density (PSD) is estimated from the noise observed in a window preceding the modification window. Given that the speech signal of the past is known exactly, and for slowly varying noise statistics, this approximation has high accuracy.

Even the relatively simple form of (3), however, is mathematically inconvenient. A work-around for this problem is found by using the asymptotically equivalent weighted distortion measure:

$$D_1 = \frac{(y+n)^2}{x} D_0 = \frac{1}{x}(y + n - xy')^2. \tag{4}$$

The value of the weighting factor is high for both very low and high signal power levels, leading to distortion emphasis in these ranges. The modified criterion, used as the basis for recovering spectral dynamics throughout this work, becomes:

$$\eta_1 = \int_\alpha^\beta \left(D_1 + \lambda n \frac{y}{x}\right) f_X(x)\, dx. \tag{5}$$

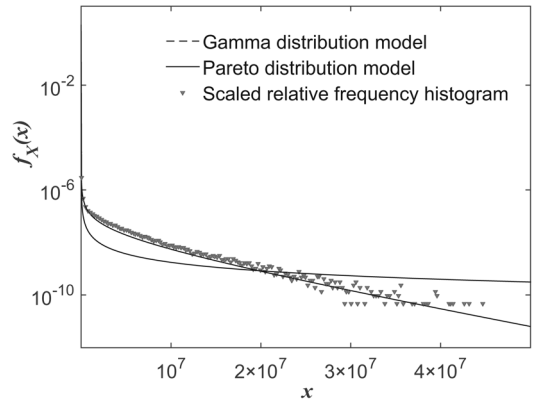


Fig. 2. Pareto distribution fit for the $(630, 870)$ Hz band.

The presence of the first order derivative in (5) suggests use of the calculus of variations [30] as the method of optimization. The possibility for finding a closed-form analytic solution to this optimization problem depends on the choice of the density function $f_X(x)$. Commonly, the Gamma distribution is used to model speech power spectral coefficients, e.g., [31]. However, exponential family distributions [32] do not lead to a tractable analysis. The two-parameter Pareto distribution [33]is adopted instead. It is a power law distribution with PDF:

$$f_X(x) = \frac{b\alpha^b}{x^{1+b}}, x \in [\alpha, \infty) \tag{6}$$

where b (shape parameter) reflects the rate of decay of the distribution tail and α is the lower bound of the support interval of x. Fig. 2 compares the modeling capacity of the Pareto and the Gamma distributions, with respect to the empirical distribution, for a particular frequency band. Instantaneous power is approximated by frame-based estimates. The large band-power values are the result of using $[-(2^{15}-1), 2^{15}]$ as the waveform's dynamic range. Maximum likelihood (ML) parameter estimates were obtained using recordings of the speech utterances from [34]. The Gamma distribution clearly stands out as a better model.

The power-law decay of the density function implies that the expectation operator over $f_X(x)$ can be evaluated for $b > 1$ only. In practice, however, estimation of the shape parameter can produce any value $b > 0$. Bounding the optimal spectral dynamics recovery range by β allows us to compute the expectation for any value of b.

An alternative to introducing an upper bound on the optimal operating range is to work with a truncated Pareto distribution [35]. Truncating the distribution, however, renders any values outside of the support infeasible and requires that all such occurrences be handled as exceptions. By using a distribution that is not bounded from above, all band-power realizations are handled with the same theoretical framework.

2) General Solution to the Single-Band Optimization Problem: Let $L(x, y, y')$ represent the integrand of (5) and $A = b\alpha^b$. Omitting the explicit dependence on x gives:

$$L(x, y, y') = \left(\frac{1}{x}(y + n - xy')^2 + \lambda n \frac{y}{x}\right) A x^{-(1+b)}. \tag{7}$$

The fundamental lemma of calculus of variations [36] provides the Euler-Lagrange equation as a necessary condition for an extremum:

$$\frac{\partial L}{\partial y} = \frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right). \tag{8}$$

Substituting the required derivatives:

$$\frac{\partial L}{\partial y} = 2Ax^{-2-b}\left(y + n - xy'\right) + \lambda Anx^{-2-b} \tag{9}$$

$$\frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right) = 2A\left(\frac{1+b}{x^{2+b}}\left(y + n - xy'\right) + x^{-b}y''\right) \tag{10}$$

into (8) and simplifying yields the following linear non-homogeneous ordinary differential equation (ODE) [37]:

$$y'' - \frac{b}{x}y' + \frac{b}{x^2}y + \frac{2bn - \lambda n}{2x^2} = 0. \tag{11}$$

The general solution to (11) is of the form:

$$y = c_1 y_{h_1} + c_2 y_{h_2} + y_{pnh}, \tag{12}$$

where y_{h_1} and y_{h_2} are two solutions to the homogeneous ODE, y_{pnh} is a particular solution to the non-homogeneous equation, and c_1 and c_2 are constants.

To find y_{h_1}, y_{h_2} and y_{pnh} we proceed as follows. One solution to the homogeneous equation is first identified. It is easily verified that x^b is a solution, i.e., $y_{h_1} = x^b$. Using the reduction of order technique, $y_{h_2} = \frac{x}{1-b}$ is found. The variation of parameters method is used to obtain $y_{pnh} = \frac{\lambda n - 2bn}{2b}$ and identify the complete solution as:

$$y = c_1 x^b + c_2 \frac{x}{1-b} + \frac{\lambda n - 2bn}{2b}. \tag{13}$$

Appendix A provides the proof that (13) is a minimizer of (5).

3) Initial and Final Conditions: To establish a particular mapping between the input power x and the output power y, it is necessary to set the constants c_1 and c_2. This is achieved by associating physically-meaningful initial and final conditions to the ODE. It is of practical interest to ensure that the mapping $y(x)$ achieves two particular points in the operating range:

$$y\left(\alpha\right) = \alpha, \tag{14}$$

$$y\left(\beta\right) = \gamma. \tag{15}$$

Solving the linear system of equations (14) and (15) gives:

$$c_1 = \frac{1}{\alpha^b\beta - \alpha\beta^b}\left(\alpha\left(\beta - \gamma\right) + \frac{\lambda n - 2bn}{2b}\left(\alpha - \beta\right)\right) \tag{16}$$

$$c_2 = \frac{1-b}{\alpha^b\beta - \alpha\beta^b}\left(\alpha^b\gamma - \alpha\beta^b + \frac{\lambda n - 2bn}{2b}\left(\beta^b - \alpha^b\right)\right). \tag{17}$$

4) Parametrization of the Final Condition: The output power γ at the upper bound $x = \beta$ of the optimal operating range can be substituted for a constant, e.g.,

$$\gamma = \beta, \tag{18}$$

or parameterized in terms of the Lagrange multiplier λ, e.g.,

$$\gamma = v\left(1 - \frac{\lambda}{2b}\right)\beta, \tag{19}$$

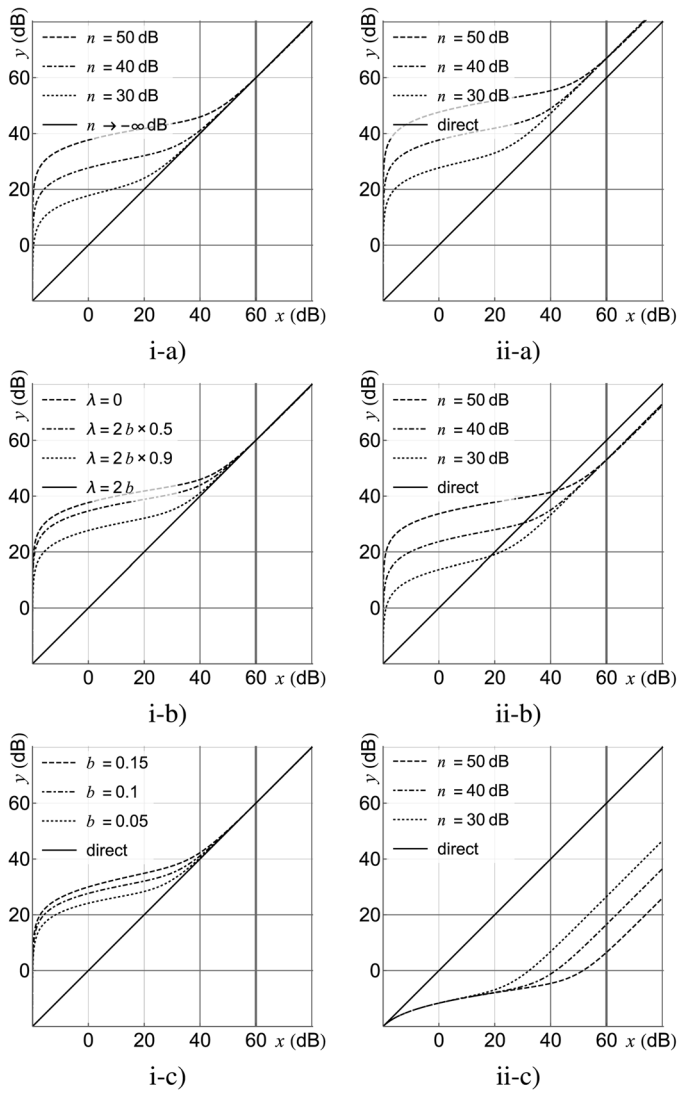


Fig. 3. Power mappings for fixed (i) and adaptive (ii) γ at: (i-a) varied noise power n for $b = 0.1$ and $\lambda = \frac{9}{10}2b$ (i-b) varied penalty λ for $n = 40$ dB and $b = 0.1$ (i-c) varied shape parameter b for $\lambda = \frac{9}{10}2b$ and $n = 40$ dB (ii-a) varied noise power n for $b = 0.1$ and penalty $\lambda = 0$ (ii-b) varied noise power n for $b = 0.1$ and $\lambda = \frac{19}{20}\lambda_u$ (ii-c) varied noise power n for $b = 0.1$ and $\lambda = \lambda_u$, where λ_u is the maximum monotonicity-preserving penalty and $\beta = 60$ dB is indicated with a thick vertical line.

where v limits the output power at $x = \beta$. The advantage of (19) over (18) is that an adaptive γ extends the range of accessible modifications. The linear parametrization in (19) is chosen for mathematical convenience. An increase in the Lagrange multiplier λ leads to a decrease in the value of γ, which is consistent with λ penalizing the power gain. Normalization by $2b$ ensures scale consistency with the remaining appearances of λ in equation (13). Note that, in general, use of boundary condition (19) does not preserve the signal in the absence of noise because $c_1 = 0$ and $c_2 = 1 - b$ do not hold.

The behavior of the mapping $y(x)$, considering both fixed and adaptive γ, is illustrated in Fig. 3. A reference level of one is used to express signal power in dB. The three plots in the left column show changes in the mapping caused by varying the values of the individual parameters n, λ and b under (18). The three plots in the right column capture changes in the mapping caused by varying noise level n for three different values of λ

under (19). The power mapping functions conform with three fundamental principles at lower power levels: i) larger noise power n increases the output power, ii) larger b (higher band-power concentration at low power levels) increases the output power and iii) larger λ decreases the output power.

Note that the shapes of the mapping curves comply with the expected behavior due to the weighting of the distortion measure in (4). For low signal powers the slope of the mapping curve is steep. A similar behavior is observed for high signal powers. The slope is moderate in the mid-power range reflecting lower sensitivity to dynamics distortion.

It is required that $y(x)$ is monotonic as non-monotonicity contradicts the objective of dynamics recovery. Monotonicity is not guaranteed for an arbitrary value of λ. We constrain the derivative of the mapping function to ensure monotonic behavior and derive a condition of the form $\lambda \in [\lambda_l, \lambda_u]$. The derivation and the expressions corresponding to case (18) and (19), are given in Appendices B and C respectively.

B. Multi-Band Dynamics Recovery

Multi-band speech modification based on (13) is discussed in this section. Increasing the spectral resolution allows for better adaptation to the statistics of the speech and the noise signals. We formulate a multi-band optimization problem in Section II-B1 and prove, in Section II-B2, that its solution ensures feed-through behavior in the absence of noise.

1) Optimization Problem: We formulate an optimization problem to distribute the available power among all bands with the objective of global dynamics recovery. Individual bands are considered of equal importance. The solution is constrained to produce monotonic mappings in all bands. The problem formulation is unaffected by the choice for γ, i.e., (18) or (19), in individual bands.

Provided that smaller $\lambda_i, i = 1 \dots I$, where i is a band index and I is the number of bands, leads to lower dynamics distortion, we minimize the sum of the squares of the normalized penalty magnitudes:

$$\underset{\lambda_i, i=1 \dots I}{\operatorname{argmin}} \sum_{i=1}^{I} \left(\frac{\lambda_i}{\lambda_{u,i}} \right)^2, \tag{20}$$

where $\lambda_{u,i}$ is the band-specific λ_u. Use of a squared error criterion equalizes dynamics importance in different bands. Different speech band-power statistics and noise band-power levels in different bands lead to $\lambda_{u,i} \neq \lambda_{u,j}, i \neq j$. This biases the objective in favor of bands with large $\lambda_{u,i}, i = 1 \dots I$. Normalization by $\lambda_{u,i}, i = 1 \dots I$ eliminates this effect.

The search region for the optimal $\lambda_i, i = 1 \dots I$ is constrained to a range where monotonicity is guaranteed by:

$$\lambda_{l,i} \leq \lambda_i < \lambda_{u,i}, i = 1 \dots I, \tag{21}$$

where $\lambda_{l,i}$ is the band-specific λ_l. Power preservation is enforced, on average, through the constraint:

$$\sum_{i=1}^{I} \mathrm{E}_{X_i} \left[y_i \left(x_i | \lambda_i \right) \right] = \sum_{i=1}^{I} \mathrm{E}_{X_i} \left[x_i \right], x_i \in [\alpha_i, \beta_i], \tag{22}$$

where expectation is taken over the only random variable according to the adopted model. Both the constant and the penalty-dependent $\gamma_i, i = 1 \dots I$ lead to an affine dependence on $\lambda_i, i = 1 \dots I$ in the left-hand-side of (22):

$$E_{X_i} \left[y_i \left(x_i | \lambda_i \right) \right] = f_i + m_i \lambda_i, x_i \in [\alpha_i, \beta_i], i = 1 \dots I, \tag{23}$$

where f_i and m_i are given by the expressions in Appendix D. The multi-band optimization problem is a convex separable quadratic program (QP), which can be solved efficiently [38].

2) Feed-Through Behavior in Noise-Free Conditions: It is required that in the absence of noise, the speech signal remains unchanged. We next show that the proposed method satisfies this constraint.

Let the parametrization (19) be used in all bands and assume that v is band-independent. In the absence of noise, the maximum penalty $\lambda_{u,i}$ becomes:

$$\lambda_{u,i} = 2 b_i \frac{v - 1}{v}, i = 1 \dots I. \tag{24}$$

When $\lambda_i = \lambda_{u,i}$ and $n_i = 0, i = 1 \dots I$, the band-specific constants $c_{1,i}$ and $c_{2,i}$ simplify to:

$$c_{1,i} = 0, \tag{25}$$

$$c_{2,i} = 1 - b_i. \tag{26}$$

Substituting (25), (26) and $n_i = 0, i = 1 \dots I$ into (13) and simplifying, gives:

$$y_i = x_i, i = 1 \dots I. \tag{27}$$

Any penalty $\lambda_i < \lambda_{u,i}, i = 1 \dots I$ leads to an increase in the output power for $x_i \in [\alpha_i, \beta_i]$. The power preservation constraint (22) together with (27) enforce $\lambda_i = \lambda_{u,i}, i = 1 \dots I$ as the optimal solution.

The case of having $\gamma_i = \beta_i$ in one of the bands is considered next. It is readily shown, starting from (13), that for $\gamma_i = \beta_i$ and independently of the value of the corresponding penalty λ_i, feed-through behavior occurs when $n_i = 0$. It follows that a combination of fixed and λ-dependent $\gamma_i, i = 1 \dots I$ preserves the signal in the absence of noise and under a power preservation constraint.

III. Practical Considerations

Aspects related to the effective operation of a multi-band speech modification system are discussed next. An overarching paradigm is that instantaneous power is approximated with estimates derived from overlapping frames extracted from the input speech signal. The estimation of the shape and scale parameters b_i and $\alpha_i, i = 1 \dots I$ from the statistical models of the speech band-powers and $\beta_i, i = 1 \dots I$ is addressed in Section III-A. Perceptual adjustment to account for using marginal distributions for the speech band-powers is considered in Section III-B. Setting the v parameter from (19) is discussed in Section III-C. The speech modification algorithm used in the evaluation of the spectral dynamics recovery approach is summarized in Section III-D.

A. Estimation of parameters b_i, α_i and β_i

The parameters b_i, α_i and $\beta_i, i = 1 \dots I$ need to be estimated prior to application of the method. The ML estimators for b_i and α_i from the PDF of the Pareto distribution are [39]:

$$\hat{\alpha}_i = \min\left\{x_{i,1}, \cdots, x_{i,J}\right\}, i = 1 \dots I, \tag{28}$$

and

$$\hat{b}_i = J\left(\sum_{j=1}^{J} \log\left(x_{i,j}\right) - J\log\left(\hat{\alpha}_i\right)\right)^{-1}, i = 1 \dots I, \tag{29}$$

where J is the number of available data points. It is readily verified that the shape parameter is scale-invariant.

The upper bound $\beta_i, i = 1 \dots I$ of the range where optimal speech modification takes place, is set to the highest observed band-power in the training database:

$$\hat{\beta}_i = \max\left\{x_{i,1}, \cdots, x_{i,J}\right\}, i = 1 \dots I. \tag{30}$$

This was a design decision reflecting the form of the estimator for the lower bound of the target range.

The parameters α_i and $\beta_i, i = 1 \dots I$ are not scale-invariant. A change in the long-term variance of the input signal is accounted for by scaling $\hat{\alpha}_i$ and $\hat{\beta}_i$ with the ratio of the current variance and the reference variance.

B. Solution Biasing

The system described thus-far does not use any knowledge of the perception of speech apart from the hypothesized importance of the spectral dynamics. In general, it amplifies the individual frequency bands of speech to alleviate the influence of the noise on the dynamics. Use of marginal distributions for the band-power statistics increases the possibility for amplifying speech components that are not perceived in a quiet environment. In practice, this effect is strongest for the high-frequency components of voiced speech.

In the current system we prevent over-amplification of inaudible speech components by biasing the modified signal spectrum towards natural speech. As stronger biasing is needed for voiced speech, a rudimentary low-complexity voicing classifier is implemented. Classification is based on the value of the first-order correlation coefficient r_1 of the speech frame. A biasing weight w_k with perceptual motivation is assigned to each class. Biasing is performed in the log-domain to avoid drastic changes in the output power distribution based on:

$$\tilde{p}_l = \frac{p_l^{1-w_k} q_l^{w_k}}{\sum_{i=1}^{I} p_i^{1-w_k} q_i^{w_k}}, w_k \in [0,1], l = 1 \dots I, \tag{31}$$

where $\mathrm{p} = \{p_1, \cdots, p_I\}$ and $\mathrm{q} = \{q_1, \cdots, q_I\}$ are the frame-specific band-power distributions of the modified (prior to biasing) and the natural speech frame respectively. The frame index was omitted for succinct notation. The denominator in (31) ensures that the frame power before and after the biasing procedure is exactly the same. The output power distribution becomes identical to that of natural speech for $w_k = 1$, whereas $w_k = 0$ prevents biasing. Note that as the noise level decreases, the output band-power distribution becomes more similar to that of the input signal. Thus, in the absence of noise, the distribution biasing procedure does not alter the signal.

In practice, $K = 4$ is sufficient for smooth transition between frames. The weights were tuned subjectively, and in the absence of noise. The objective of reducing distortions while maximally preserving the character of the modified speech resulted in:

$$\begin{aligned} &w_1 = 0.6, r_1 \in [0.99, 1), k = 1, \\ &w_2 = 0.5, r_1 \in [0.97, 0.99), k = 2, \\ &w_3 = 0.4, r_1 \in [0.95, 0.97), k = 3, \\ &w_4 = 0.3, r_1 \in (0, 0.95), k = 4. \end{aligned} \tag{32}$$

The SDR method extended by biasing of the power distributions is referred to as biased-SDR (bSDR). Obtaining the bSDR from the SDR optimal solution can be regarded as adaptive post-filtering.

C. Setting the v Parameter

The parameter v of (19) affects the maximum output power at the upper end of the optimal operating range when an adaptive γ (eq. (19)) is used, and is related to the monotonicity of the mapping curves. Modes are eliminated from the optimal mapping range for $\lambda_i \in [\lambda_{l,i}, \lambda_{u,i}]$ when $v \in [\max(v_{l,i}, v_{u,i}), \infty)$ is satisfied. The expressions for $v_{l,i}$ and $v_{u,i}$ and the analysis are provided in Appendix C.

In Section II-B it was assumed that v is band-independent. It is likely that higher intelligibility results can be achieved by introducing band-dependence but this track is not pursued. We set v to the closest integer value, which ensures that:

$$v \geq \max\left(v_{l,i}, v_{u,i}\right), i = 1 \dots I \tag{33}$$

is satisfied in at least 99% of the cases encountered in the training database for three different noise types at -9 dB SNR. In practice, bands in which (33) cannot be satisfied will use a fixed-γ (eq. (18)) mapping.

The particular noise types and SNR value are used in the listening test performed to assess the intelligibility of modified speech. This SNR level was used as the most severe condition for stationary noise in the Public Hurricane Challenge [40].

D. Algorithm

The algorithm for spectral dynamics recovery is presented in this section. A block diagram is given in Fig. 4. The upright notation $\mathtt{x}, \mathtt{n}$ and $\mathtt{y}$ denotes the respective waveforms. Band indexes are not included for simplicity. It is understood that $\alpha, \beta, \gamma, \lambda, b, x, n$ and y represent the corresponding sets of parameters. ζ indicates whether the mapping assigned to each band uses a fixed (eq. (18)) or an adaptive γ (eq. (19)).

Three background gray levels are used to show the frequency of execution of different operations. The lightest background represents operations performed once per frame. A darker background is used for the less frequent operations that follow an update in the noise statistics. In practice, the period T of these operations corresponds to the duration of an utterance (approximately 2.5 s). The darkest background is used for an operation performed once only.

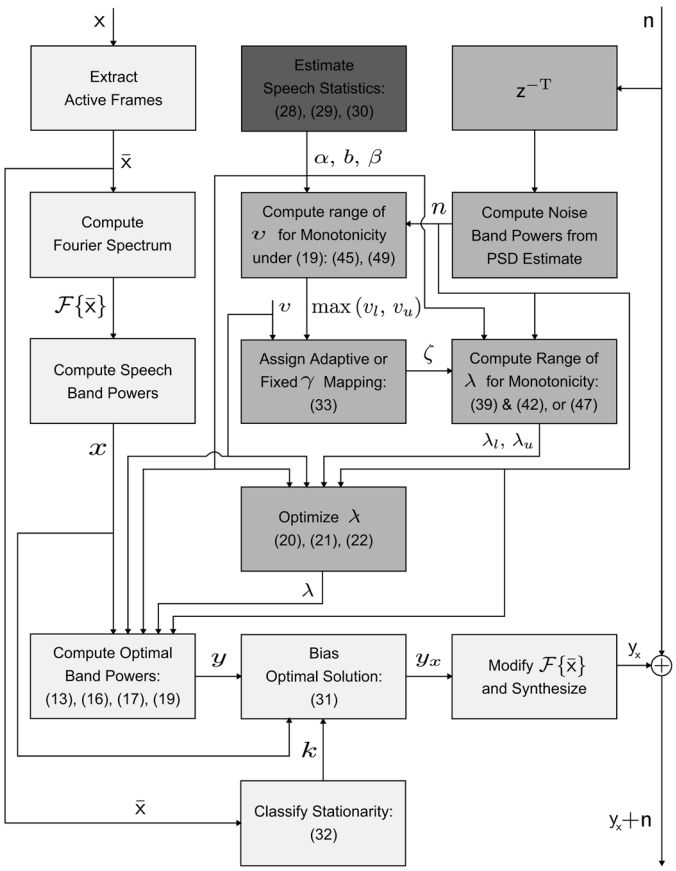


Fig. 4. Step-by-step listing of the proposed algorithm. Numbers in brackets are references to equations. The background gray levels reflect the operation execution frequency as follows: light - once per frame, medium - once per noise statistics update and dark - once.

The following two operations are not shown explicitly in Fig. 4. Rescaling of α and β (discussed in Section III-A) takes place when the long-term variance of the input signal changes. Abrupt changes in the spectral gain function are eliminated by convolution with a smooth (Hann) window. This operation reduces artifacts in the modified speech.

The algorithm attempts to assign adaptive-γ (eq. (19)) mappings to as many of the bands as possible. The motivation, discussed previously, is that use of equation (19) offers more flexibility for signal power distribution over equation (18). For bands where the *a priori* choice of υ leads to a violation of the monotonicity constraint (under eq. (19)), a fixed γ is used. Once each band is assigned a mapping function, the multi-band optimization for λ is performed.

The frame duration reflects the requirement for quasi-stationarity in the extracted segments. The frame update rate presents a trade-off between complexity and degree of smoothing in the processed speech. The speech band-powers are computed from the periodogram of individual frames extracted from the clean signal. The bands are non-overlapping (to simplify the optimization) and uniformly spaced on a Mel scale [27]. The noise band-powers are estimated from the history of the noise by averaging the periodogram over the frames within an analysis window of predefined duration.

IV. Method Evaluation

Results from the validation of the proposed approach to modifying speech for improved intelligibility in noise are presented in this section. System-related settings and complexity analysis are given in Section IV-A. The reference system used in the subjective evaluation of the proposed method is introduced in Section IV-B. Objective evaluation results are presented in Section IV-C followed by subjective evaluation results in Section IV-D.

A. System Settings and Complexity

$I = 14$ bands uniformly distributed on a Mel scale [27] in the range $[0.1, 7.8]$kHz provide sufficient level of detail in the power spectral domain while incurring a relatively low computational complexity. A 2048–long FFT is used for a frame length of 416 samples at a sampling rate of 16 kHz. A frame overlap of 50% was used, resulting in an algorithmic delay of 13 ms. A 50–sample long Hann window is convolved with the band-gain factors (at the spectral bin level) to smooth transitions between adjacent bands.

The estimates for b_i, α_i and $\beta_i, i = 1 \ldots I$ were obtained using British English recordings of the Harvard Sentences [34] from [40]. The 720 sentences were variance-equalized prior to parameter estimation. The shape parameter values were:

$$\begin{aligned}\mathbf{b} = 10^{-2} \cdot \{&6.7\ 7.8\ 7.4\ 8.2\ 9.0\ 8.8\ 8.7 \\ &9.6\ 9.5\ 9.9\ 11.9\ 14.9\ 15.4\ 16.9\},\end{aligned} \tag{34}$$

indicating a trend of increase in the concentration of band-powers towards lower power levels with an increase in the band center frequency. The variance threshold for detecting speech active frames was set to 50 dB below that of the maximum observed frame power. Changes in this threshold affect the values $\alpha_i, i = 1 \ldots I$ and, consequently, the mapping functions. This threshold level allows signal re-synthesis without audible discontinuities.

Following the procedure identified in Section III-C, $\upsilon = 5$ was used over all bands and test conditions. Noise power spectral density estimates were obtained by averaging the periodograms of frames extracted from the 500 ms of noise signal preceding the information-bearing speech signal in the recordings for each noise-utterance pair.

The complexity of the proposed and the reference methods was evaluated in Matlab as the average over 135 utterances with a mean duration of 2 seconds. The $quadprog(\cdot)$ routine was used in combination with the active-set optimization algorithm (under default settings) to solve the multi-band problem. A MacBook A1466 computer with an i7 1.7 GHz processor allowed for an execution 15 times faster than real-time operation and five times faster than the reference method [16] in Matlab [41]. A DELL 6520 computer with an i7 2.2 GHz processor achieved roughly the same processing time for the proposed method while reducing the processing time taken by the reference method by half.

B. Reference System

SII-optimally-modified speech [16] was used as the reference during the subjective evaluation of the dynamics recovery

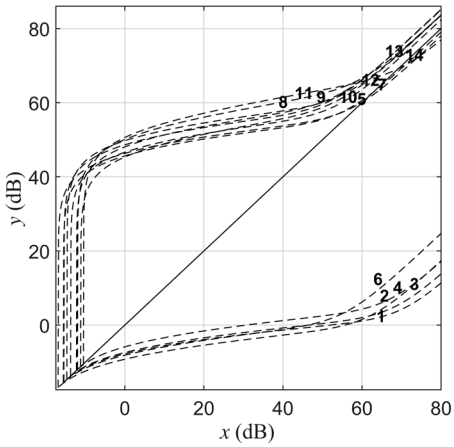


Fig. 5. Band-specific input-output power maps for −9 dB SNR speech-shaped noise in a system with 14 bands.

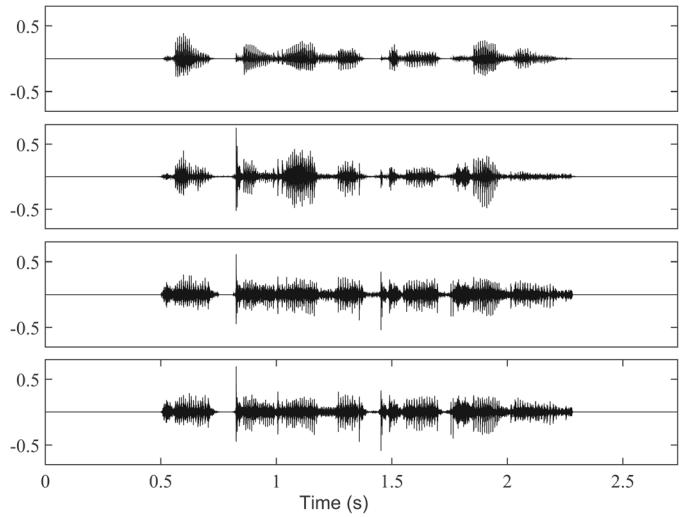


Fig. 6. Speech waveforms (for presentation at −9 dB SNR speech-shaped noise) for (in order from top) natural, SII-modified, SDR-modified and bSDR-modified speech.

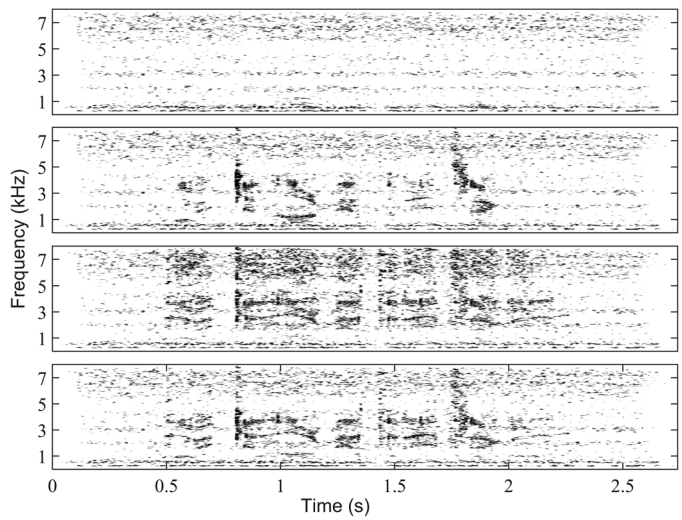


Fig. 7. Noisy (−9 dBSNR speech-shaped noise) speech spectrograms for (in order from top) natural, SII-modified, SDR-modified and bSDR-modified speech.

approach. SII [3] is a well-established model for evaluating the intelligibility of speech in noisy environments. This method is of particular interest as speech is modified based on perceptual considerations represented by band-importance weights, and therefore different from SDR.

Two adjustments are made to the original method from [16] with the intention of addressing an on-line communication scenario. On the one hand, the noise PSD estimate is obtained from the 500 ms of noise signal preceding the utterance. On the other, the optimal filter computation is based on the history of the speech signal as opposed to the actual utterance being processed. The availability of the future signal cannot be assumed for on-line communication scenarios. This is addressed by using the preceding sentence, which is of a similar duration and uttered by the same speaker.

C. Objective Validation

The behavior of the proposed method is first illustrated in terms of the time and spectral domain effects it induces. We consider a fourteen-band system configuration and a test condition of speech-shaped noise at −9 dB SNR.

The optimal mapping functions for all fourteen bands are illustrated in Fig. 5. The power is reduced in a number of bands (typically bands with low shape parameter values) in favor of the remaining bands. Band five, unlike its adjacent bands four and six is boosted. This is explained in view of the higher value of the shape parameter b_5, cf. eq. (34), and the higher noise level compared to band six.

The signal waveforms of the natural and the modified signals, including the reference, SDR and bSDR methods, are presented in Fig. 6. All three speech modification techniques enhance weaker transient parts of the signal. It is in the spectral domain that the differences between the three methods become more apparent. Fig. 7 shows the noisy spectrograms for the natural and the modified speech signals from Fig. 6. While SII-optimally-modified speech produces a number of clear islands in the noisy signal spectrogram, it lacks the continuity obtained with the two methods based on recovering the spectral dynamics. The cost of dynamics recovery are relatively lower output power levels.

The spectrograms also provide an insight into the differences between SDR and bSDR. High-frequency power, which does not identify any clear trends in the evolution of the spectral dynamics in the unbiased case, is reallocated in favor of emphasizing the more pronounced dynamics at the lower frequency range.

Objective evaluation results for three noise types using the SII implementation from the reference system are presented in Fig. 8. To simplify the presentation, improvement over natural speech is shown. Each value is an average over 135 utterances from the test database [42]. All three modification techniques achieve higher intelligibility than natural speech. Note that SII does not consistently rate biasing as an enhancement. The objective results suggest higher performance in most of the considered range for SDR and bSDR in speech-shaped and babble noise. In factory noise the proposed method is expected to perform slightly worse than the reference method. Typical noise PSD estimates are presented for completeness in Fig. 9.

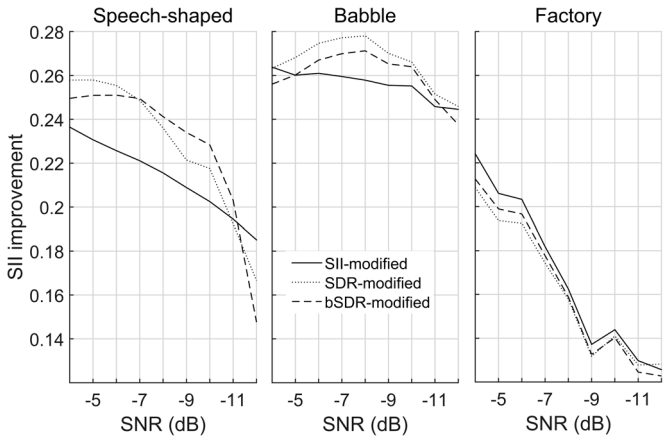


Fig. 8. Intelligibility improvement over natural speech evaluated based on the SII implementation of the reference system.

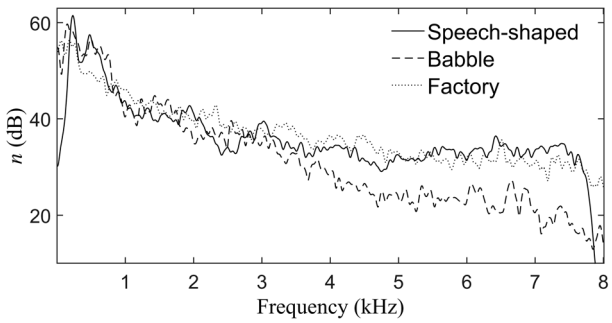


Fig. 9. Average periodograms of the three noise types from the test conditions.

D. Subjective Validation

The results from the subjective evaluation of the proposed approach are presented in this section. The protocol for the listening test is given in Section IV-D1 and the corresponding test results are given in Section IV-D2.

1) Listening Test Set-Up: To facilitate subject recruitment, a Swedish sentence database [42] was used. It contains 180 sentences divided into phonetically-balanced sets of 15 sentences. Only the first 135 sentences were used during testing to ensure a test duration preventing listener fatigue. The test database is different from the training database, which ensures compliance with a practical application scenario where the data is not known *a priori*.

The recordings from the chosen database were preprocessed by a noise suppression algorithm [43] to remove stationary recording noise. After normalizing the variance of each utterance, a silence period of 500 ms was added before and after to form the clean test signals. Noisy signal mixtures were then obtained using speech-shaped, multi-speaker babble and factory noise [44] at -9 dB SNR. The computation of the SNR excludes the silence periods before and after the utterance. The modified speech from both the proposed and the reference algorithms was renormalized to -9 dB SNR to compensate for power deviations due to the short utterance duration and the inaccuracy of the adopted statistical model.

Nine native listeners, aged 25 - 35 and without known auditory impairments, were recruited and compensated financially for their participation. The protocol can be summarized as follows. The test is computer-based and diotic presentation

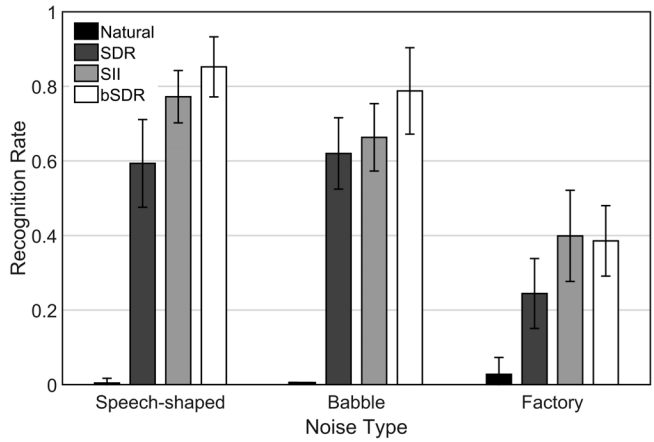


Fig. 10. Mean subjective intelligibility scores and ± 1 standard deviation bars at -9 dB SNR.

is done through a pair of Beyerdynamic DT 770 headphones. Before the test begins, six sentences (not included in the test) are played back to allow the subject to adjust the presentation level. Thereafter, using a particular combination of processing method, noise type and sentence set, a subject is presented with a sequence of sentences. After a single presentation of a sentence, the subject is prompted to input their version of it using a text-based computer interface. Upon completing the task and pressing "Enter" the test continues with the presentation of the next sentence. Over the nine subjects all possible combinations of sentence sets, noise types and processing methods are covered. In addition, the order of presentation of the processing methods is shifted cyclically after a set of three subjects to ensure grounds for fair comparison. On average 25 minutes were sufficient for completing the test.

Joint testing of the three speech modification methods and natural speech requires use of all 180 utterances from the test database and extends significantly the duration of the listening test. For this reason, natural speech was excluded from the test when the three processing methods were compared. The results for the intelligibility of natural speech in noise, presented in Section IV-D2, were taken from a previous test run comparing the intelligibility of natural speech, SII-optimally-modified speech and speech modified with an earlier version of SDR using exactly the same test protocol but different listeners.

Intelligibility scores were computed for each sentence as the ratio of the number of correctly recognized and the total number of words. Per-sentence scores were then averaged over sentence sets and used for comparison and significance analysis, c.f., Section IV-D2.

2) Subjective Evaluation Results: The results from the listening test performed using the test protocol described in Section IV-D1 are presented in this section. Score averages and standard deviations over subjects are shown in Fig. 10. p-values obtained with the Wilcoxon signed-rank test [45] are given in Table I. This test assesses the difference in the mean ranks of two data sets without assuming a particular probability density function. Commonly, p-values smaller than 0.05 are considered to indicate that the difference is significant.

All of the included modification techniques improve intelligibility significantly. SDR achieves lower performance

TABLE I
WILCOXON SIGNED-RANK TEST SIGNIFICANCE LEVELS

SSN	Nat.	SDR	SII	bSDR
Nat.	1	0.004	0.004	0.004
SDR	-	1	0.004	0.004
SII	-	-	1	0.078
bSDR	-	-	-	1

BBL	Nat.	SDR	SII	bSDR
Nat.	1	0.004	0.004	0.004
SDR	-	1	0.598	0.004
SII	-	-	1	0.008
bSDR	-	-	-	1

FRY	Nat.	SDR	SII	bSDR
Nat.	1	0.004	0.004	0.004
SDR	-	1	0.008	0.016
SII	-	-	1	0.715
bSDR	-	-	-	1

compared to the reference and bSDR systems for reasons discussed in Section III-B. bSDR outperforms the reference system in speech-shaped and multi-speaker babble noise while achieving insignificantly lower performance in factory noise. The low recognition scores for natural speech are likely caused by the relatively short duration of the test sentences (four words on average) leading to low predictability of the test material and reduced listener adaptation time.

A comparison of the subjective and the objective evaluation results shows that SII generally overrates the performance of SDR. The advantage of bSDR over the reference system in speech-shaped and babble noise is captured by the model. Similarly, for factory noise the objective model identifies correctly the modest advantage of the reference method.

V. CONCLUSIONS

Spectral dynamics recovery is an effective approach to improving speech intelligibility in noise. Calculus of variations provides the tools to optimize a distortion criterion quantifying the deviation of the band-specific power dynamics of the noisy from that of the clean speech. The solution is a mapping from natural to optimally-modified speech band-powers that adapts to the statistics of both the speech and the noise. The band-specific power mapping provides the basis for designing a speech modification system with low computational complexity and algorithmic delay. Future work will focus on sophisticating the multi-band optimization criterion and removing the need for post-processing of the optimal solution.

APPENDIX A
VERIFYING THAT OPTIMAL y IS A MINIMIZER

The Euler-Lagrange equation is a necessary condition for an optimizer and does not reveal if its solution, the stationary point $y^*(x)$, is extremal and whether it is a minimizer or a maximizer. Checking the second order (Legendre) necessary condition we verify that y^* can be a minimizer but not a maximizer [30]:

$$\frac{\partial^2 L\left(x,y,y'\right)}{\left(\partial y'\right)^2} = 2b\alpha^b x^{-b} > 0, y \equiv y^*, x \in (\alpha,\beta), \quad (35)$$

where α, β and b are all positive.

We next look at a sufficient condition for the stationary point y^* to be a minimizer [30]. The objective is to show that $L(x,y,y')$ is jointly convex in (y,y'). This is achieved by proving that the following Hessian is positive semidefinite:

$$H_L = \begin{bmatrix} \frac{\partial^2 L}{(\partial y)^2} & \frac{\partial^2 L}{\partial y \partial y'} \\ \frac{\partial^2 L}{\partial y' \partial y} & \frac{\partial^2 L}{(\partial y')^2} \end{bmatrix} = \frac{2b\alpha^b}{x^{2+b}} \begin{bmatrix} 1 & -x \\ -x & x^2 \end{bmatrix}. \quad (36)$$

As the eigenvalues of H_L

$$\operatorname{eig}\{H_L\} = \begin{cases} 0 \\ 2b\alpha^b \frac{x^2+1}{x^{2+b}} \end{cases} \quad (37)$$

are non-negative on $x \in (\alpha,\beta)$, y^* is a minimizer.

APPENDIX B
MONOTONICITY OF $y(x)$ FOR $\gamma = \beta$

Let λ_l and λ_u define the lower and the upper bound on a range of λ values that preserve the monotonicity of $y(x)$, $x \in [\alpha,\beta]$. Considering the fixed γ case, we derive these bounds starting with constraints on the derivative of the mapping. The relation $b < 1$, which ensures that only a single extremal point can be achieved, is used for the purpose.

The condition:

$$y'\left(x = \alpha | \alpha,\beta,b,\lambda,n\right) \geq 0, \quad (38)$$

ensures that $y(\alpha)$ is non-decreasing. Solving for λ gives:

$$\lambda \leq \lambda_u \equiv 2b\left(1 + \frac{\alpha^b\beta - \alpha\beta^b}{n\xi}\right), \quad (39)$$

Where

$$\xi = \alpha^{b-1}\beta b - \beta^b - \alpha^b b + \alpha^b. \quad (40)$$

The lower bound λ_l is obtained similarly, by ensuring that $y(x)$ is non-decreasing at $x = \beta$:

$$y'\left(x = \beta | \alpha,\beta,b,\lambda,n\right) \geq 0. \quad (41)$$

Considering that $\lambda \geq 0$ must hold, we obtain:

$$\lambda \geq \lambda_l \equiv \max\left(0, 2b\left(1 + \frac{\alpha^b\beta - \alpha\beta^b}{n\kappa}\right)\right), \quad (42)$$

where

$$\kappa = -\alpha\beta^{b-1}b + \alpha^b + \beta^b b - \beta^b. \quad (43)$$

APPENDIX C
MONOTONICITY OF $y(x)$ FOR $\gamma = v(1 - \frac{\lambda}{2b})\beta$

The adaptive nature of γ complicates the monotonicity analysis compared to the case with constant γ discussed in Appendix B. The presence of the two adjustable parameters λ and v in (19) suggest a two-step approach: the λ-s that ensures a particular behavior for $y(x)$ are first identified. The feasible range for v is then derived.

We require that the derivative $y'(x)$ is non-negative at $x = \beta$ for $\lambda_l \equiv 0$, which is the minimum allowed penalty:

$$y'\left(x = \beta | \alpha,\beta,b,\lambda_l,v,n\right) \geq 0, \lambda_l = 0. \quad (44)$$

This constraint eliminates a maximum on $x \in (\alpha, \beta)$. It is readily derived that the inequality is satisfied when:

$$v \geq v_l \equiv \frac{\alpha\beta^b(1-b) + n(1-b)\beta^b + \alpha\beta^{b-1}bn - \alpha^b n}{\alpha^b\beta - \alpha\beta^b b}. \quad (45)$$

Following a similar approach, the possibility for a minimum on $x \in (\alpha, \beta)$ is eliminated. We impose the constraint:

$$y'(x = \alpha|\alpha, \beta, b, \lambda_u, v, n) = 1, \quad (46)$$

which is different from (38) (used for fixed γ). This choice ensures that the input signal is not modified in the absence of noise for a multi-band signal modification scenario, cf. Section II-B2. Solving eq. (46) for λ_u gives:

$$\lambda_u = 2b\frac{n\xi + \alpha^b\beta(b-1)(1-v)}{n\xi - \alpha^b\beta(b-1)v}. \quad (47)$$

The range of v where $y'(x = \beta) \geq 0$ for $\lambda = \lambda_u$ is identified next. Constraining the function to be increasing at $x = \alpha$ and non-decreasing at $x = \beta$ ensures that it is monotonically-increasing on $x \in [\alpha, \beta)$ (when $b < 1$ holds). The constraint:

$$y'(x = \beta|\alpha, \beta, b, \lambda_u, v, n) \geq 0, \beta > \alpha \quad (48)$$

leads to:

$$v \geq v_u \equiv n\frac{-\alpha^b\beta\kappa + \alpha\beta^b\xi}{\alpha^b\beta(\alpha^b\beta - \alpha\beta^b)}. \quad (49)$$

To summarize the above analysis, $y(x)$, $x \in [\alpha, \beta]$ is monotonically increasing for

$$\lambda \in \left[0, 2b\frac{n\xi + \alpha^b\beta(b-1)(1-v)}{n\xi - \alpha^b\beta(b-1)v}\right] \quad (50)$$

when

$$v \geq \max(v_l, v_u). \quad (51)$$

APPENDIX D
EXPRESSIONS FOR f AND m FROM EQ. (23)

The band index i is omitted for succinct notation. The expressions for f and m under (18) are:

$$f = \frac{n(\beta-\alpha)}{\alpha^b\beta - \alpha\beta^b}b\alpha^b\log\left(\frac{\beta}{\alpha}\right) - n\varphi + \frac{\alpha^b(n+\beta) - \beta^b(n+\alpha)}{(\alpha^b\beta - \alpha\beta^b)(1-b)}b\alpha^b\left(\beta^{-b+1} - \alpha^{-b+1}\right) \quad (52)$$

and

$$m = \frac{n(\alpha-\beta)}{2(\alpha^b\beta - \alpha\beta^b)}\alpha^b\log\left(\frac{\beta}{\alpha}\right) + \frac{n}{2b}\varphi + \frac{n\left(\beta^b - \alpha^b\right)}{2(\alpha^b\beta - \alpha\beta^b)(1-b)}\alpha^b\left(\beta^{-b+1} - \alpha^{-b+1}\right), \quad (53)$$

where

$$\varphi = 1 - \left(\frac{\alpha}{\beta}\right)^b. \quad (54)$$

The expressions for f and m under (19) become:

$$f = \frac{n(\beta-\alpha) + \alpha\beta(1-v)}{\alpha^b\beta - \alpha\beta^b}b\alpha^b\log\left(\frac{\beta}{\alpha}\right) - n\varphi + \frac{\alpha^b(n+\beta v) - \beta^b(n+\alpha)}{(\alpha^b\beta - \alpha\beta^b)(1-b)}b\alpha^b\left(\beta^{-b+1} - \alpha^{-b+1}\right) \quad (55)$$

and

$$m = \frac{n(\alpha-\beta) + \alpha\beta v}{2(\alpha^b\beta - \alpha\beta^b)}\alpha^b\log\left(\frac{\beta}{\alpha}\right) + \frac{n}{2b}\varphi + \frac{n\left(\beta^b - \alpha^b\right) - \alpha^b\beta v}{2(\alpha^b\beta - \alpha\beta^b)(1-b)}\alpha^b\left(\beta^{-b+1} - \alpha^{-b+1}\right). \quad (56)$$

REFERENCES

[1] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[2] J. B. Allen, "How do humans process and recognize speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[3] American National Standard, Methods for the Calculation of the Speech Intelligibility Index, 1997.

[4] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, pp. 331–348, 2003.

[5] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Amer.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[6] A. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[7] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.

[8] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 510–524, Sep. 1999.

[9] The Listening Talker Project. [Online]. Available: http://listening-talker.org

[10] B. A. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 1, pp. 22–32, Mar. 1969.

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

[12] T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012.

[13] 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. Signal Process. Audio Acoust. Workshop*, 2005, pp. 62–65.

[14] P. S. Chanda and S. Park, "Speech intelligibility enhancement using tunable equalization filter," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, 2007, pp. 613–616.

[15] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. Eur. Signal Process. Conf*, 2010, pp. 1919–1923.

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

[17] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, pp. 493–496.

[18] 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, Signal Process*, 2012, pp. 4061–4064.

[19] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, 2012.

[20] C. V. 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," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 665–686, 2013.
[21] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.
[22] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 1035–1045, May 2013.
[23] 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, U.K.: Cambridge University Engineering Department, 2009.
[24] E. M. Z. Golumbic, D. Poeppel, and C. E. Schroeder, "Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective," *Brain and Lang.*, vol. 122, pp. 151–161, 2012.
[25] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neurosci.*, vol. 15, pp. 511–517, 2012.
[26] P. N. Petkov and W. B. Kleijn, "Preservation of speech spectral dynamics enhances intelligibility," in *Proc. Interspeech*, 2013.
[27] B. C. Moore, *An Introduction to the Psychology of Hearing*. Amsterdam, The Netherlands: Elsevier, 2004.
[28] O. Ghitza, "Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm," *Front. Psychol.*, vol. 2, 2011.
[29] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *J. Neurophysiol.*, vol. 107, pp. 78–89, 2012.
[30] B. Chachuat, *Nonlinear and Dynamic Optmization: From Theory to Practice*. Lausanne, Switzerland: Automatic Control Laboratory, EPFL, 2007.
[31] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-gaussian priors," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 253–256, Mar. 2013.
[32] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*. Chichester, U.K.: Wiley, 1978.
[33] B. C. Arnold, *Pareto Distributions*. Fairland, MD, USA: International Cooperative Publishing House, 1983.
[34] , "IEEE recommnded practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.
[35] I. B. Aban, M. M. Meerschaert, and A. K. Panorska, "Parameter estimation for the truncated Pareto distribution," *J. Amer. Stat. Assoc.*, vol. 101, no. 473, pp. 270–277, 2006.
[36] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. New York, NY, USA: Wiley, 1953, vol. 1.
[37] M. Tenenbaum and H. Pollard, *Ordinary Differential Equations: An Elementary Textbook for Students of Mathematics, Engineering, and the Sciences*. New York, NY, USA: Harper & Row, 1963.
[38] J.-P. Dussault, J. A. Ferland, and B. Lemaire, "Convex quadratic programming with one constraint and bounded variables," *J. Math. Progr. Amer.*, vol. 36, no. 1, pp. 90–104, 1986.
[39] R. E. Quandt, "Old and new methods of estimation and the Pareto distribution," *Metrika*, vol. 10, no. 1, pp. 55–82, 1966.
[40] M. Cooke, C. Mayo, C. V. Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Commun.*, vol. 55, pp. 572–585, 2013.
[41] [Online]. Available: http://www.mathworks.com,Matlab2013a, 2013
[42] J. Beskow, L. Cerrato, B. Granström, D. House, M. Nordstrand, and G. Svanfeldt, "The swedish PF-Star multimodal corpora," in *Proc. LREC Workshop Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, 2004.
[43] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.
[44] 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.
[45] D. F. Bauer, "Constructing confidence sets using rank statistics," *J. Amer. Stat. Assoc.*, vol. 67, pp. 687–690, 1972.

Petko N. Petkov holds a B.Sc. in communication engineering from the Technical University of Sofia and an M.Sc. and a Ph.D. in electrical engineering from KTH Royal Institute of Technology, Stockholm, Sweden. He was a Research and Development Engineer with Global IP Solutions in 2006–2007. He is currently a Research Engineer with the Speech Technology Group, Cambridge Research Laboratory, Toshiba Research Europe Limited. His research interests include the application of signal processing and machine learning to problems in speech and audio processing.

W. Bastiaan Kleijn has been a Professor at Victoria University of Wellington since 2010. He is also a Professor at Delft University of Technology (part-time) and was a Professor at KTH, where he headed 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. He holds a Ph.D. in electrical engineering from Delft University of Technology and an M.S.E.E. from Stanford. He also earned a Ph.D. in soil science and an M.S. in physic from the University of California, Riverside. He is a Fellow of the IEEE.