Preservation of Speech Spectral Dynamics Enhances Intelligibility

Petko N. Petkov\textsuperscript{1}, W. Bastiaan Kleijn\textsuperscript{1,2}

\textsuperscript{1} School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm, Sweden
\textsuperscript{2} School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
petkov@kth.se, bastiaan.kleijn@ecs.vuw.ac.nz

Abstract

We propose a method for the enhancement of intelligibility in scenarios where speech is rendered in a noisy environment. The method is based on the hypothesis that intelligibility is a monotonic function of the degree of preservation of the speech spectral dynamics. The accuracy of the speech spectral dynamics can then be traded against the power of the rendered speech signal. We can either maximize the dynamics accuracy given the signal power, or minimize the signal power given the dynamics accuracy. In our implementation, the spectral dynamics is quantified as the difference of the mel cepstra between time frames of the speech signal. We compared the speech rendered by our implementation against both natural speech and a reference method, for the scenario where signal power is minimized given a target dynamics accuracy, and observed a significantly improved intelligibility. The low system delay, and the low complexity and memory requirements make the new method particularly suitable for real-time applications.

Index Terms: speech intelligibility, spectral dynamics

1. Introduction

Speech modification for improved intelligibility in noise is an active research topic. Various methods have been proposed to address, in particular, the scenario of additive noise. These can be classified into rule-based and objective-measure-based methods. Use of an objective measure does not guarantee higher performance \cite{1}. It provides, however, a more general framework in which different modifications can be applied to the speech signal \cite{2}, and the possibility to monitor their effectiveness through a figure of merit.

Commonly, the methods that optimize an objective measure of intelligibility do so under an energy limitation constraint \cite{2, 3, 4, 5} to prevent excessive output levels. As a result, such methods provide the optimal modification parameters for the available energy but do not ensure the sufficiency of this energy to reach a certain intelligibility level. In theory, it is possible to identify the energy required to achieve an \textit{a-priori} specified figure of merit with any objective measure. The practical challenges that arise are related to i) identifying the mapping between the figure of merit and intelligibility and ii) obtaining the optimal, among several solutions, e.g., in a band-based modification system where multiple solutions lead to the same figure of merit.

We propose a method for energy-adjusting speech modification using an indicator of intelligibility motivated from recent studies in auditory perception, e.g., \cite{6, 7, 8, 9}. These emphasize the importance of the temporal evolution of the speech signal to intelligibility. To introduce multiple degrees of freedom and to facilitate the use of well-established concepts in models of the auditory periphery \cite{10, 11} we use a band-based approach. Spectral dynamics preservation (SDP) was, thus, chosen as a representative name for the method.

SDP represents a fundamentally different approach to speech rendering compared to other measure-based methods. It does not aim to recover a local spectro-temporal relationship between the signal and the noise (or the noisy signal), e.g., in terms of SNR \cite{3, 5}, correlation \cite{4}, or visibility (glimpses) \cite{12}. Instead, it targets the preservation, up to an \textit{a-priori} selected level, of a time-evolving relationship in the channels of an auditory filter-bank.

We derive the conditions for SDP in terms of the differences of mel frequency cepstral coefficients (MFCCs) \cite{13}. MFCCs are commonly used features for automatic speech recognition, e.g., \cite{14}. They are perceptually motivated and, therefore, attractive for application to speech rendering. By choosing a target SDP level, a trade-off is established between intelligibility recovery and energy usage. The SDP level maps to band-specific power thresholds, above which the dynamics of the band-limited signals are preserved. Below these thresholds preservation is not possible and the signal is left unaltered.

Limited subjective evaluation of the intelligibility enhancement capabilities of SDP was performed. As a reference, we used a recently proposed method \cite{4} based on its state-of-art performance and capability to operate at a short system delay. The reference method optimizes a perceptual distortion measure and provides the possibility for comparing the perceptual attributes of the output speech from the two enhancement strategies. The evaluation results favor the proposed method and motivate further and more extensive investigation of the effect of spectral dynamics on intelligibility.

The remainder of this paper is organized as follows. The conditions for achieving SDP and the formulation of the optimization problem for the speech modification parameters are presented in Section 2. Implementation-related considerations are provided in Section 3. Experimental validation results are shown in Section 4 followed by conclusions.

2. Theoretical foundations

The conditions for preservation of the spectral dynamics are defined in Section 2.1. Optimization of the speech modification parameters is discussed in Section 2.2, where we focus on minimizing the power required to achieve a target SDP level.

2.1. Conditions for preserving spectral dynamics

The SDP conditions for additive noise are derived here for a band-specific gain-adapting signal modification. Spectral dynamics are represented by differences in MFCC features, commonly referred to as deltas \cite{14} in an automatic speech recognition (ASR) front-end.
To introduce the necessary notation we first recall the sequence of operations for computing the MFCCs from a windowed signal frame [13]:

1. \( \Phi_{x_j} = X_j \circ x_j \), where \( j \) is a frame index and \( x_j \in \mathbb{R}^N \) is a random vector of signal samples with Fourier transform \( X_j \); computes the periodogram;
2. \( \Psi_j = M^T \Phi_{x_j} \), maps the spectrum powers to the Mel scale using a bank of overlapping, triangular filters (the columns of \( M \));
3. \( s_{i,j} = \log (m_i^T \Phi_{s_j}) \), computes the log of the Mel-scale band-power in band \( i = 1..I \) and frame \( j \);
4. \( A_j = C^T S_j \), \( S_j = [s_{i,j} \cdots s_{I,j}]^T \), decorrelates the log band-powers using the DCT transform (matrix \( C \)).

Note that, since the DCT is a unitary transform,

\[ S_j = C A_j, \]  

(1)

Let the difference between the MFCCs of two frames be

\[ \Delta_{A_j} = A_{j+i} - A_j, \]  

(2)

where the frame offset \( i \) is an arbitrary integer, i.e., the analysis is not constrained to successive frames. The difference between the log band-powers of the respective frames is:

\[ \Delta_{s_i} = S_{j+i} - S_j. \]  

(3)

From equations (1), (2) and (3) it follows that

\[ \Delta_{s_i} = C \Delta_{A_j}, \]  

(4)

which establishes that preserving \( \Delta_{s_i} \) is a sufficient condition for preserving \( \Delta_{A_j} \).

In the presence of additive and uncorrelated noise, the Mel-scale band-powers \( \Psi_j \) take the form:

\[ \Psi_j = M^T \Phi_{x_j} + M^T \Phi_{n_j}, \]  

(5)

where \( \Phi_{n_j} \) is the noise periodogram associated with frame \( j \). The log band-power differences are:

\[ \Delta_{s_i,j} = \log \left( \frac{m_i^T \Phi_{s_{j+i}} + m_i^T \Phi_{n_{j+i}}}{m_i^T \Phi_{s_j} + m_i^T \Phi_{n_j}} \right). \]  

(6)

Our objective is to modify the speech signal such that the original log band-power differences are preserved. To facilitate the analysis, we replace the overlapping triangular filters in the columns of \( M \) by non-overlapping filters. To reduce the dependence of the result on specific auditory modelling assumptions we ignore the effect of frequency tuning associated with the triangular filter shape and work with uniform response filters.

Allowing for adjustment of the signal band-powers using scaling coefficients \( b_{i,j} \) and \( b_{i,j+i} \), respectively, we observe that preservation is achieved when

\[ \frac{b_{i,j+i} m_i^T \Phi_{s_{j+i}} + m_i^T \Phi_{n_{j+i}}}{b_{i,j} m_i^T \Phi_{s_j} + m_i^T \Phi_{n_j}} = \frac{c_i m_i^T \Phi_{s_{j+i}}}{c_i m_i^T \Phi_{s_j}}, \]  

(7)

for any value of the coefficient \( c_i \). Equation (7) has two unknowns and, therefore, multiple solutions. We obtain a particular solution by establishing, separately, the equality of the numerators and the denominators from the left and the right-hand side of (7).

Importantly, note that untying \( b_{i,j} \) from \( b_{i,j+i} \) allows us to avoid large system delay (represented in frames by \( l \)). Further, it leads to a simple expression of the form:

\[ b_{i,j+i} = c_i \frac{m_i^T \Phi_{s_{j+i}}}{m_i^T \Phi_{s_j}}, \]  

(8)

which enables computationally-efficient speech modification.

The coefficients \( b_{i,j+i}, \forall l \) operate in the power domain and must be non-negative. We note that the parameters \( c_i, i = 1..I \) determine the trade-off between SDP preservation accuracy and power consumption. Increasing \( c_i \) allows the preservation of the dynamics, for band \( i \), in regions with lower signal levels. The cost is an increase in the output signal power.

When SDP, in a spectral band, cannot be achieved, due to signal band-powers falling below the preservation threshold, an adjustment to the modification coefficients is required. One possibility is to mute the band. The absence of a measure of signal audibility, in the current formulation of the method, implies that muting may lead to audible distortion. We choose to leave the pass-band signal unaltered when SDP cannot be achieved. To avoid gain inconsistencies in transitions from SDP-infeasible to SDP-feasible frames, we further constrain \( b_{i,j} \), \( \forall i, j \) to:

\[ b_{i,j} \geq 1, \forall i, j. \]  

(9)

In Section 2.2 we address the derivation of \( c_i, i = 1..I \) with a perspective on achieving an average (over all bands) target SDP level with the minimum output power.

### 2.2. Modification parameter optimization for SDP

We next derive the optimal coefficients \( c_i, i = 1..I \) for which, on average, a target SDP level is achieved by scaling the frame-based signal gain in each band by \( \sqrt{b_{i,j}} \). We assume the statistical properties of natural speech, with fixed long-term variance, to be known and define the random variables

\[ w_i = m_i^T \Phi_{s_i}, i = 1..I \]  

(10)

where the short-term speech periodogram \( \Phi_{s_i} \) is no longer restricted to a particular frame. The marginal probability and cumulative density functions of these random variables are denoted by \( p(w_i) \) and \( P(w_i), i = 1..I \).

Similarly, we denote the random noise band-powers by

\[ v_i = m_i^T \Phi_{n_i}, i = 1..I. \]  

(11)

In practice, however, we will work with the deterministic estimates \( \hat{V}_i \) of \( v_i, i = 1..I \) under the assumption of slowly-varying noise statistics.

Using the above notation, we express \( c_i, i = 1..I \) as:

\[ c_i = 1 + \frac{\hat{V}_i}{w_i} = 1..I \]  

(12)

where \( w_i \) is the threshold band-power, above which SDP is achieved. The formulation in equation (12) was chosen as it ensures that for \( w_i = w_i^\# \), the gain becomes unity as indicated by the updated form of equation (8):

\[ b_{i,j} = 1 + \frac{\hat{V}_i}{w_i^\#} - \frac{\hat{V}_i}{w_i}, i = 1..I. \]  

(13)

The average output power of the enhanced signal is:

\[ \hat{\Phi} = \sum_{i=1}^{I} \left[ \frac{w_i^\#}{w_i} p(w_i) dw_i + \int_{w_i}^{\infty} b_i w_i p(w_i) dw_i \right]. \]  

(14)
We define the SDP level \( r \) as the average over the contributions of the individual bands:
\[
    r = \frac{1}{I} \sum_{i=1}^{I} \left[ 1 - P \left( w_i^* \right) \right].
\]
(15)

This formulation provides a degree of freedom that we use to identify the most power-efficient combination of the optimal levels \( w_i^* \), \( i = 1..I \). We achieve this by minimizing the average output power subject to the SDP level constraint:
\[
    w_i^*, \; i = 1..I = \arg \min_{w_i^*} \Phi,
\]
subject to
\[
    \frac{1}{I} \sum_{i=1}^{I} \left[ 1 - P \left( w_i^* \right) \right] = r.
\]

None of the commonly used probability density functions with positive support allows for a closed-form solution to (16). In general, it is not possible to make a claim for the convexity of the Lagrangian [15] of (16). However, both the objective function and the constraint are monotonically increasing in \( w_i^*, \; i = 1..I \), suggesting that poor local optima are not a concern. In Section 3 we present an efficient approximate solution to (16).

### 3. Implementation

A greedy optimization algorithm for solving (16) is presented in this section. The solution is approximate as it is obtained through the quantization of the variables \( w_i, \; i = 1..I \). We replace \( p(w_i) \) and \( P(w_i) \), \( i = 1..I \) by the corresponding probability mass and cumulative density functions \( g(\tilde{w}_i) \) and \( G(\tilde{w}_i) \), \( i = 1..I \). The optimization problem transforms to:
\[
    k_i^*, \; i = 1..I = \arg \min_{k_i^*} \tilde{\Phi},
\]
subject to
\[
    \frac{1}{I} \sum_{i=1}^{I} \left[ 1 - G \left( \tilde{w}_i,k_i^* \right) \right] \geq r,
\]
where
\[
    \tilde{\Phi} = \sum_{i=1}^{I} \left[ k_i^*-1 \sum_{k_{i-1}}^{K_i} \tilde{w}_{i,k_i} g \left( \tilde{w}_{i,k_i} \right) + \sum_{k_{i-1}}^{K_i} \tilde{b}_i \tilde{w}_{i,k_i} g \left( \tilde{w}_{i,k_i} \right) \right],
\]
(18)

\( k_i \) is an index over the possible realizations of the discrete random variable \( \tilde{w}_i \), and \( \tilde{b}_i \) and \( \tilde{c}_i \) are given by:
\[
    \tilde{b}_i = \tilde{c}_i = 1 + \frac{V_i}{\tilde{w}_{i,k_i^*}}.
\]
(19)

The optimization algorithm increases the SDP level by applying a single-step decrement of the threshold \( \tilde{w}_{i,k_i^*} \) in the band that leads to the minimum increase in output power. After a finite number of threshold decrements, the SDP constraint is satisfied and the optimal thresholds \( \tilde{w}_{i,k_i^*}, \; i = 1..I \) are identified. The algorithm can be summarized as follows:

1. Set \( k_i^* = K_i, \; i = 1..I \) and compute \( \tilde{\Phi} \);
2. Compute the output power increase due to the decrement \( k_i^* = K_i - 1, \; i = 1..I \) separately for each band;
3. Select the band \( q \) for which the output power increase is minimal and update \( k_q^* = k_q^* - 1 \);
4. Compute the expected power increase from further lowering of the threshold in band \( q \);
5. Compute the current SDP level;
6. Exit if the SDP constraint is satisfied or loop to step 3.

The convergence of the algorithm is guaranteed by the monotonically increasing SDP level with a decrease of the threshold in any spectral band. Assuming that the number \( K_i \) of possible realizations of \( \tilde{w}_i \) is the same in all bands, i.e., \( K_i = K, \; i = 1..I \), convergence is attained with \( r/K \) iterations on average.

For practical reasons, the separation between the realizations of \( \tilde{w}_i, \; i = 1..I \) is not uniform. The approximating probability mass functions are first derived from histograms of the log band-powers and then mapped to the linear domain.

### 4. Experimental Results

The experimental configuration and the results from the objective and the subjective evaluation of the proposed method for speech rendering in noise are presented in this section. We used a set-up with \( I = 32 \) bands linearly-spaced on a Mel scale in the range \([100, 7500]\) Hz, corresponding to the spectral range of the signals in the development database. The sampling frequency was 16 kHz. An analysis frame length of 12 ms and update length of 10 ms were used. The analysis window was tapered in the overlap regions corresponding to the front and the tail of a Hanning window. Prior to spectral analysis zero-padding to 2048 samples was performed. The band-power densities were approximated using \( K = 100 \) levels.

To estimate the band-power statistics of natural speech we used 160 sentences from the Harvard sentence database [16] (sets 46 through 61). The recordings were taken from [17]. Two types of additive noise were considered: speech-shaped (SSN) and multi-speaker babble (BBL) [18]. The a-priori SNR in the presented experiments was \(-6\) dB. A single estimate of the noise power spectral density (PSD) was obtained, in the desired resolution, for use in the rendering algorithm from the first 500 ms of the noise recording assigned to each test sentence. In practice the noise PSD estimate can be updated arbitrarily often.

The reference method (PD) [4] was set to operate at a system delay of one frame, where the frame length was 32 ms and the update length was 16 ms. As the SDP method modifies the SNR, the power of the natural speech signal is equalized to the output power of the (SDP) modified signal to ensure fair comparison. Similarly, the reference system output is obtained from processing the re-scaled natural speech signal.

The optimal coefficients \( c_i, \; i = 1..I \) at \(-6\) dB SNR for SSN and BBL are shown in Figure 1. An SDP level \( r = 0.35 \) provided an attractive trade-off between intelligibility improvement and signal power increase. Both noise types lead to optimal contours of similar shape. The three peaks observed in the range \([1500, 4500]\) Hz are likely the effect of the speech formants. The resonances lead to higher speech power and, consequently, cheaper recovery of the dynamics. As expected, the outcome changes depending on the noise properties. The increase in both coefficient series towards the high end of the spectral range is caused by a decrease in the noise PSD estimate, which leads to the possibility for power-efficient SDP recovery.

The effect on the signal is illustrated in Figure 2 for SSN using sentence one from set 47 of [16]. The variances of both the natural and the modified speech signals are equal. We observe a characteristic contour adjustment of the temporal envelope of the full-band signal. The noisy spectrograms corresponding to the natural and the modified signals are shown in Figure 3. The output SNR for this particular example was 0.5 dB.

Limited subjective evaluation of the proposed algorithm in a comparison to natural speech and speech processed by the reference system was performed using six non-native English speakers. The test took place in a silent room, us-
ing a pair of Beyerdynamic DT 770 headphones and a rudimentary text-based computer interface. Each subject screened six sets (47 through 52) of sentences in total, for a different combination of test condition and processing methods. Following a single presentation of each sentence the subject was prompted to enter their version of the sentence. The recognition rates at the sentence level were computed as the ratio of correctly identified keywords (all words except ‘a’, ‘the’, ‘in’, ‘to’, ‘on’, ‘is’, ‘and’, ‘of’, ‘for’, ‘at’) and the total number of keywords. The per-set averaged raw recognition scores are presented in Table 1.

As the output SNR of the individual sentences varies (due to their short duration) we performed the comparison of the subjective recognition scores on a per-set basis as seen from Table 2. To mitigate the effect of the inconsistent language proficiency of the subjects, we normalized the scores for each subject by their best per-set recognition score. The significance analysis of the normalized scores is shown in Table 3, using the conservative Wilcoxon signed rank test [19]. These results indicate the significance of the intelligibility improvement over natural speech (Nat.) and the speech processed by the reference method (SSN condition only) even with this low number of subjects.

5. Conclusions

Based on psycho-acoustic results it can be hypothesized that preserving the spectral dynamics of speech in the context of additive noise leads to significantly enhanced intelligibility. Our results confirmed this hypothesis. The low computational complexity and the low delay of the proposed method make it suitable for time-sensitive applications. SDP offers the flexibility of operating in one of two modes: minimizing the output signal power under a constraint on the spectral dynamics recovery level or maximizing the spectral dynamics recovery level under a constraint on the output signal power.

6. Acknowledgements

The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230.

The authors are grateful to Mengqiu Zhang for a stimulating discussion on the topic of preserving spectral dynamics.
7. References


