SIDECHAIN HARMONIC ENHANCEMENT OF NOISE CORRUPTED SPEECH FOR HEARING IMPAIRED LISTENERS

Kamil Wojcicki\textsuperscript{1}, Kelly Fitz\textsuperscript{1}, Karrie Recker\textsuperscript{1}, Don Reynolds\textsuperscript{2}, Tao Zhang\textsuperscript{1}

\textsuperscript{1}Starkey Hearing Technologies, Eden Prairie, MN 55344, United States
\textsuperscript{2}Starkey Hearing Research Center, Berkeley, CA 94704, United States
E-mail: \{kamil_wojcicki, kelly_fitz, karrie_recker, don_reynolds, tao_zhang\}@starkey.com

ABSTRACT

This work presents a single channel speech enhancement approach aimed at improving speech clarity for hearing impaired listeners under challenging listening conditions. The proposed method applies nonlinear distortions to speech components isolated from the observed noisy signal using aggressive speech enhancement. The enhanced components are then mixed back into the noisy signal. The results show that the proposed approach significantly improves speech clarity in noise.

Index Terms—speech enhancement, speech clarity, harmonic distortion, hearing aids

1. INTRODUCTION

Enhancement of speech corrupted by noise is one of the biggest challenges for the hearing aid industry. One problem shared by conventional noise reduction (NR) algorithms, e.g., \cite{1, 2, 3}, is that they do not improve the local signal-to-noise ratio (SNR) within individual time-frequency units. Typically, attenuation is applied to individual time-frequency units according to a continuous gain function based on an estimate of the local SNR. As such, conventional approaches scale both speech and noise components in a given time-frequency unit by the same amount. For this reason, the local SNR within a given unit remains unchanged after processing. Thus, while speech quality or listening comfort may be improved, speech intelligibility is typically degraded, or at best unchanged \cite{4, 5}. Unlike the conventional approaches which focus solely on attenuation of the unwanted noise from the noisy mixture, the focus of the present work is on addition of new speech-related information. The new content is based on spectral regions characterized by high local SNR and is mixed into spectral regions with potentially low local SNR.

In commercial music production, harmonic enhancement (HE) is used frequently as a “sweetening” technique \cite{6}. It is itself a distortion process, and is generally only applied in small amounts, to prevent the “sweetening” from being perceived as objectional distortion or corruption of the signal \cite{7}. The end goal is to make certain frequency ranges more apparent, clearer, and more “pleasing” to listen to. Typically, harmonics are generated by applying nonlinear distortion to the music, or to the individual vocals or instruments, possibly with band-pass filtering of the signal before and/or after the nonlinearity \cite{8}. The use of nonlinear distortions also finds application in the area of artificial bandwidth extension \cite{9}. For example, speech signals transmitted over public switched telephone networks are band limited to 300–3400 Hz. The bandwidth and quality of band-limited speech signals can be improved through diligent application of nonlinear distortions at the receiving end \cite{9}.

While such perceptually desirable effects can be achieved when applying harmonic enhancement to the target material (whether it be speech or music), it is generally undesirable to introduce harmonically related components of the noise (or the mixture). For this reason, the use of harmonic enhancement is typically limited to relatively noise-free signals. In order to overcome this, we propose to apply aggressive speech enhancement (ASE) as a preprocessor to harmonic enhancement. In this way, only the parts of the target-plus-masker mixture strongly dominated by the target speech are subjected to nonlinear distortions. This processing takes place in a sidechain, and the newly generated speech-like information is then added back into the time-frequency units in the main signal path. As such, the proposed approach represents a unique combination of speech enhancement and signal enhancement techniques. In this work, we focus specifically on single channel enhancement of speech corrupted by additive noise at challenging mixture SNRs (i.e., $\leq 5$ dB), with the primary objective of improving speech clarity for hearing impaired listeners under such conditions.

The remainder of this work is organized as follows. Details of the proposed algorithm are presented in Section 2. Experiments used for algorithm parameter selection are discussed in Section 3. Section 4 presents speech enhancement experiments used to evaluate the performance of the proposed method in terms of speech clarity and sound quality. Results and discussion are presented in Section 5. Conclusions are given in Section 6.

2. SIDECHAIN HARMONIC ENHANCEMENT

A block diagram of the proposed sidechain harmonic enhancement (SHE) algorithm is given in Fig. 1. The main signal path, along with a sidechain, are shown. In the main path, NR is optionally applied, with its spectral gain function hard limited to $G_{\text{max}}$, i.e., the gain function is permitted to apply at most $G_{\text{max}}$ dB attenuation. Thus, NR is enabled when $G_{\text{max}} > 0$, and disabled when $G_{\text{max}} = 0$. Only very mild NR is permitted (i.e., $G_{\text{max}} \leq 3$) in order to avoid addition of perceptually-distracting NR-related artifacts and to ensure any unwanted SHE-introduced artifacts are masked.

In the sidechain, the noisy signal is bandlimited at the ASE input to between $f_1$ and 7500 Hz. Speech-dominated components are then isolated from the bandlimited signal in the ASE block. This can be achieved using binary mask types of approaches, e.g., \cite{10, 11, 12, 13}, or by utilizing more traditional speech enhancement methods, e.g., \cite{14, 1, 3, 15}, tuned to achieve aggressive noise reduction. HE is then applied in order to generate harmonically-related speech-like components in other parts of the spectrum. For this purpose, a nonlinear transformation can be applied to the time domain signal, e.g., soft-clipping, cubic compression, or half-wave
rectification. Alternatively, HE could also be achieved directly in the
frequency domain, e.g., by frequency transposition. The output sig-

main from HE is then bandlimited to between $f_2$ and 7500 Hz. The
enhanced signal is computed as the sum of this bandlimited signal
scaled by $g_2$, the ASE output signal scaled by $g_1$, and the main path
signal. In the above $f_1$ and $f_2$ are tunable frequency parameters, and
$G_{\text{max}}$, $g_1$, and $g_2$, are tunable gain parameters.

The bandlimiting operations at the input of ASE and at the out-
put of HE serve the following purpose. The former selects the spec-
tral region to be subjected to harmonic enhancement, while the lat-
ter selects the range of harmonically enhanced frequencies to be re-
tained for mixing with the main path signal. The output of ASE can
also be mixed back in directly with the noisy signal in the main path,
with the gains $g_1$ and $g_2$ controlling how much of ASE and HE pro-
cessed signals are added back into the main path. While the former
contributes (adds energy) to spectral regions with already high SNR,
the latter hopes to contribute new information to low SNR spectral
regions.

Our realization of the SHE algorithm utilized a short-time
Fourier analysis-modification-synthesis system for frequency do-
main processing. This system was based on the weighted overlap-
add (WOLA) discrete Fourier transform filterbank [16, 17]. The
sampling frequency was set to 16 kHz. The WOLA analysis du-
ration was set to 4 ms, with a frame shift of 0.5 ms. The size of the
discrete Fourier transform (DFT) in the WOLA analysis was set to
32. The DFT was computed using a fast Fourier transform (FFT) al-
gorithm. The processing was applied in the first sixteen frames.
The energy in the Nyquist band was set to zero. Conjugate symmetry of
the spectrum was ensured prior to synthesis.

The minimum mean-square error (MMSE) short-time spectral
amplitude (STSA) estimator [1], in conjunction with the unbiased
noise power estimator proposed in [18], were used for the NR and
ASE blocks. The reference implementation given in [19] was uti-
lized within the WOLA-based system. Time constants, used for
smoothing of noise power, speech probability and the a priori SNR
estimates, were tuned so as to achieve aggressive isolation of speech
dominated spectral regions. Time domain half-wave rectification
was used for HE.

3. PARAMETER SELECTION EXPERIMENTS

The parameter space for the SHE method proposed in Section 2
was explored by hearing impaired listeners using a genetic algorithm
(GA) approach [20]. The GA is a biologically inspired optimization
routine that uses the concept of survival of the fittest, in which
the best solutions to a problem evolve while poorer solutions die off
[20]. A benefit of a GA is that it facilitates exploration of several
parameters at once to quickly determine which solutions (parama-
ter sets) produce the best results for the listener. This is especially
important in situations in which the parameters may interact with
each other, and therefore cannot be optimized individually. Rather
than testing all possible combinations of parameter settings, which
may be very time consuming, or not at all feasible, a GA explores
the complex parameter space in a directed way, reducing the amount
of time spent testing regions where the parameter space consistently
generates poor solutions. Locally optimal solutions are determined
by having a listener compare two sets of parameters, or “genes,”
from a pool of genes. Genes that are rated poorly eventually die off
(i.e., they are removed from the gene pool), whereas genes that the
listener rates highly remain in the gene pool and contribute to the fu-
ture generations of genes. New genes are created when highly-rated
genes either mutate (change some of their parameters) or “mate”
with other genes to create “child” genes.

The GA approach was chosen for this study, because there were
many parameters for which the optimal solutions were unknown.
Further, these parameters may interact with each other, and therefore
could not be optimized individually. It was also unknown whether
the optimal solution(s) would be participant-specific or general to
the group. The goal was to identify parameter sets associated with
improved speech clarity. The searched parameter space consisted of
five variables, discretized as shown in Table 1.

The GA testing consisted of hearing impaired participants (see
Section 4.1) listening to pairs of stimuli, and indicating on a discrete
scale which of the two stimuli had higher speech clarity and by how
much (slightly, moderately or strongly so). For this purpose a GA
toolbox given in [21] was utilized.

The results of GA testing are summarized in Table 2. Individ-
ual as well as “global” parameter sets are shown. For the former, the
fittest parameter set identified for each subject through GA-based lis-
tening test is listed. The latter (MILD and STRONG) were computed
based on k-means clustering of the fittest parameter sets identified
for each subject. The MILD parameter set includes noise reduction
in the main signal path with strong harmonic harmonic enhancement

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_{\text{max}}$</td>
<td>{0, 2, 3}</td>
</tr>
<tr>
<td>$g_1$</td>
<td>${-\infty, -6, 0, 6, 9.5}$</td>
</tr>
<tr>
<td>$g_2$</td>
<td>${-6, 0, 6, 9.5}$</td>
</tr>
<tr>
<td>$f_1$</td>
<td>${250, 750, 1250}$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>${250, 750, 1250, 2500, 3750}$</td>
</tr>
</tbody>
</table>

Table 1. SHE parameters along with discretized values for the GA
tests. The listed values for the gain parameters ($G_{\text{max}}$, $g_1$, and $g_2$)
are all expressed here in dB, while the frequency parameters ($f_1$ and
$f_2$) are given in Hz.
Table 2. SHE parameter sets determined based on GA tests.

<table>
<thead>
<tr>
<th>TREATMENTS</th>
<th>PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GAIN (dB)</td>
</tr>
<tr>
<td></td>
<td>(G_{\max})</td>
</tr>
<tr>
<td>GLOBAL</td>
<td>Mild</td>
</tr>
<tr>
<td>P01</td>
<td>3</td>
</tr>
<tr>
<td>P02</td>
<td>0</td>
</tr>
<tr>
<td>P03</td>
<td>3</td>
</tr>
<tr>
<td>P04</td>
<td>0</td>
</tr>
<tr>
<td>P05</td>
<td>0</td>
</tr>
<tr>
<td>P06</td>
<td>3</td>
</tr>
<tr>
<td>P07</td>
<td>2</td>
</tr>
<tr>
<td>P08</td>
<td>0</td>
</tr>
<tr>
<td>P09</td>
<td>2</td>
</tr>
<tr>
<td>P10</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2. SHE parameter sets determined based on GA tests.

<table>
<thead>
<tr>
<th>TREATMENTS</th>
<th>PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GAIN (dB)</td>
</tr>
<tr>
<td></td>
<td>(G_{\max})</td>
</tr>
<tr>
<td>GLOBAL</td>
<td>Mild</td>
</tr>
<tr>
<td>P01</td>
<td>3</td>
</tr>
<tr>
<td>P02</td>
<td>0</td>
</tr>
<tr>
<td>P03</td>
<td>3</td>
</tr>
<tr>
<td>P04</td>
<td>0</td>
</tr>
<tr>
<td>P05</td>
<td>0</td>
</tr>
<tr>
<td>P06</td>
<td>3</td>
</tr>
<tr>
<td>P07</td>
<td>2</td>
</tr>
<tr>
<td>P08</td>
<td>0</td>
</tr>
<tr>
<td>P09</td>
<td>2</td>
</tr>
<tr>
<td>P10</td>
<td>3</td>
</tr>
</tbody>
</table>

4. SPEECH ENHANCEMENT EXPERIMENTS

The performance of the proposed method was evaluated through a series of subjective tests. These tests focused on assessment of speech clarity by hearing impaired listeners [22]. Sound quality was also evaluated. The details of these experiments are given in the remainder of this section.

4.1. Subjects

Ten individuals with symmetrical mild-to-moderately severe sensorineural hearing loss (high-frequency audiometric thresholds ≤70 dB HL) participated in this study. Note that this is a common type of hearing loss observed in an aging population. The listeners were paid for their participation.

4.2. Speech and noise materials

Speech materials from the TIMIT corpus [23] were used. These were divided into practice and test sets composed of 30 and 48 sentences, respectively. Both sets were gender balanced, i.e., half of the sentences in each set belonged to female talkers and the rest belonged to male talkers. Each sentence in the test set was spoken by a different talker.

Three maskers were used, namely speech shaped noise (SSN) [24] at 0 and 5 dB mixture SNRs, as well as multi-talker babble [25] and restaurant [26] noises at -5, 0 and 5 dB mixture SNRs.

4.3. Types of stimuli

Four treatment types were investigated:

1. **Off**—the target+masker mixture (baseline)
2. **INDIVIDUAL**—individualized SHE treatment, i.e., tailored for each participant
3. **MILD**—global SHE treatment
4. **STRONG**—global SHE treatment

The latter three treatments were based on GA-derived SHE parameter sets outlined in Section 3.

4.4. Stimuli generation

The stimuli for the different treatment types were created as follows. A portion of a given masker recording was selected at random and mixed-in with a speech recording by keeping the level of the speech constant and adjusting the level of the masker to achieve a desired mixture SNR. A given type of signal processing, associated with a given treatment, was then applied—i.e., Off (no processing) or SHE processing (INDIVIDUAL, MILD or STRONG). The stimuli were pre-processed to compensate for each individual’s hearing loss. Specifically, separate prescriptions appropriate for the hearing loss in each ear were applied. The prescriptions were based on the e-STAT formula provides a mild amount of compression. Sampling frequency of 16 kHz was used for the processing. The stimuli were then up-sampled to 48 kHz, prior to being stored as two channel audio files with 24 bit pulse-code modulation encoding.

4.5. Experiment procedure

A “rate and rank” test (RRT) was used to assess speech clarity and sound quality. The RRT task allows listeners to directly compare several stimuli in each trial. Specifically, the participants were presented with a horizontal scale split into two sections. The left-hand side indicated that the treatment was considered acceptable on a given dimension, while the right-hand side indicated that the treatment was considered unacceptable on that dimension. The separate scales used to assess speech clarity and sound quality, ranged from “very clear” to “very unclear”, and from “very good” to “very poor”, respectively. Four symbols were shown above each scale. Each symbol was associated with a stimulus pre-generated as outlined in Section 4.4. The participants could listen (and re-listen) to the stimuli by clicking on the symbols. The stimuli were presented over open circumaural headphones (Sennheiser HD600, Wennebostel, Germany). Each stimulus lasted two to three seconds. The listeners were instructed to rate and rank the stimuli (associated with the four treatments) by dragging the symbols onto the continuous scale using a computer mouse. The ratings, bounded between 0 and 1, were collected for the four stimuli in each trial by the computer software.

The speech clarity and sound quality assessments were conducted over two separate sessions. At the start of each session, the participants were familiarized with the task during a short practice set composed of 6 trials. The speech materials for these trials were selected at random from the practice set (see Section 4.2). This was followed by a test that consisted of 48 trials: 2 talker genders × 2 mixture SNRs (SSN) × 3 sentences + 2 talker genders × 3 mixture SNRs × 2 maskers (babble and restaurant) × 3 sentences, i.e., each treatment was repeated 3 times for each gender, masker and mixture SNR combination. For each of the 48 trials, a unique sentence spoken by a unique talker was used. All variables (gender, masker and mixture SNR) were randomized across listeners. The treatment to symbol allocations were also randomized across trials and listeners. The listeners were encouraged to take breaks. Each session lasted approximately one hour.

4.6. Statistical model

A 5-factor, hierarchical Bayesian linear model was designed for analysis of the RRT responses. In this model, each factor (gender, masker, mixture SNR, treatment, and participant) was a predictor of the parameters of a latent variable representing the speech clarity or sound quality. This latent variable was modeled as a Gaussian whose mean was a linear combination of contributions from all
In this work, a novel speech enhancement approach was proposed and evaluated on hearing impaired listeners. The processing took place in a sidechain. The method applied nonlinear distortion to speech-dominated content isolated from the noisy mixture using aggressive form of noise reduction. The enhanced components were then mixed back into the main signal path. The results showed that the proposed approach significantly improved speech clarity. For the treatment with the largest clarity improvements, there was also an associated significant reduction in sound quality.
7. REFERENCES


