**SPEAKER DEPENDENT SPEECH ENHANCEMENT USING SINUSOIDAL MODEL**

*Pejman Mowlaee and Christian Nachbar*

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

**ABSTRACT**

In many conventional speech enhancement methods, discrete Fourier transformation is used in analysis, modification, and synthesis stages without incorporating a signal-dependent model or the prior knowledge about the underlying speaker characteristics. In this work, we integrate a sinusoidal model as speech signal model and further include speaker information captured in a trained speaker model in the form of a sinusoidal coder. We design a postfilter as a postprocessor after a conventional speech enhancement stage. We show that the proposed method significantly improves the perceived quality in particular for non-stationary noise and low signal-to-noise ratio scenarios. The improved performance predicted by instrumental metrics is further justified by subjective listening tests.

*Index Terms—* Speech enhancement, speaker-dependent model, sinusoidal model, sinusoidal coder, perceived quality.

1. INTRODUCTION

The ultimate performance of the existing conventional speech enhancement algorithms heavily relies on the accuracy of the noise power estimator as it impacts on both *a priori* and *a posteriori* SNRs required to obtain the noise suppression gain value to be applied on the noisy signal. This is a critical issue when the added noise is a non-stationary one as a biased noise estimate often leads to an inferior performance of the speech estimation. Incorporating priori information about the underlying speaker in noise or a signal-dependent model of speech signal can alleviate such limitation via decoupling the noise and speech estimators, leading to certain improvement in the perceived quality of enhanced speech possible.

Several previous attempts made steps towards incorporating a signal model or speaker model into speech enhancement framework. The authors in [1] presented an enhancement algorithm using a constrained iterative sinusoidal model as the representation of the characteristics of speech. The algorithm derived smoothed trajectories of amplitudes and frequencies and was shown to bring certain improvement in the objective speech quality measures. Other methods employed auto-regressive (AR) features as their signal model [2] or in combination with a hidden Markov model as their speaker model [3] shown to properly handle the non-stationary noise scenarios. The authors in [4] proposed to employ codebooks trained on speech and noise representing the AR-modeled spectral envelopes. Different codebooks for different noise types were trained and classified according to the observed noisy signal. The shapes of the amplitude envelopes captured in codebooks are represented by linear prediction coefficients (LPC) [5] or cepstral coefficients [6] and the appropriate gain values of the individual amplitude envelopes are obtained.

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-ICT-2011-7-288121.

In many conventional speech enhancement methods, discrete Fourier transformation is used in analysis, modification, and synthesis stages without incorporating a signal-dependent model or the prior knowledge about the underlying speaker characteristics. In this work, we integrate a sinusoidal model as speech signal model and further include speaker information captured in a trained speaker model in the form of a sinusoidal coder. We design a postfilter as a postprocessor after a conventional speech enhancement stage. We show that the proposed method significantly improves the perceived quality in particular for non-stationary noise and low signal-to-noise ratio scenarios. The improved performance predicted by instrumental metrics is further justified by subjective listening tests.

**Index Terms—** Speech enhancement, speaker-dependent model, sinusoidal model, sinusoidal coder, perceived quality.

1. INTRODUCTION

The ultimate performance of the existing conventional speech enhancement algorithms heavily relies on the accuracy of the noise power estimator as it impacts on both *a priori* and *a posteriori* SNRs required to obtain the noise suppression gain value to be applied on the noisy signal. This is a critical issue when the added noise is a non-stationary one as a biased noise estimate often leads to an inferior performance of the speech estimation. Incorporating priori information about the underlying speaker in noise or a signal-dependent model of speech signal can alleviate such limitation via decoupling the noise and speech estimators, leading to certain improvement in the perceived quality of enhanced speech possible.

Several previous attempts made steps towards incorporating a signal model or speaker model into speech enhancement framework. The authors in [1] presented an enhancement algorithm using a constrained iterative sinusoidal model as the representation of the characteristics of speech. The algorithm derived smoothed trajectories of amplitudes and frequencies and was shown to bring certain improvement in the objective speech quality measures. Other methods employed auto-regressive (AR) features as their signal model [2] or in combination with a hidden Markov model as their speaker model [3] shown to properly handle the non-stationary noise scenarios. The authors in [4] proposed to employ codebooks trained on speech and noise representing the AR-modeled spectral envelopes. Different codebooks for different noise types were trained and classified according to the observed noisy signal. The shapes of the amplitude envelopes captured in codebooks are represented by linear prediction coefficients (LPC) [5] or cepstral coefficients [6] and the appropriate gain values of the individual amplitude envelopes are obtained.

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-ICT-2011-7-288121.

**Fig. 1.** Block diagram of the proposed algorithm: speaker-dependent speech enhancement using sinusoidal model.

**ABSTRACT**

In many conventional speech enhancement methods, discrete Fourier transformation is used in analysis, modification, and synthesis stages without incorporating a signal-dependent model or the prior knowledge about the underlying speaker characteristics. In this work, we integrate a sinusoidal model as speech signal model and further include speaker information captured in a trained speaker model in the form of a sinusoidal coder. We design a postfilter as a postprocessor after a conventional speech enhancement stage. We show that the proposed method significantly improves the perceived quality in particular for non-stationary noise and low signal-to-noise ratio scenarios. The improved performance predicted by instrumental metrics is further justified by subjective listening tests.

*Index Terms—* Speech enhancement, speaker-dependent model, sinusoidal model, sinusoidal coder, perceived quality.

1. INTRODUCTION

The ultimate performance of the existing conventional speech enhancement algorithms heavily relies on the accuracy of the noise power estimator as it impacts on both *a priori* and *a posteriori* SNRs required to obtain the noise suppression gain value to be applied on the noisy signal. This is a critical issue when the added noise is a non-stationary one as a biased noise estimate often leads to an inferior performance of the speech estimation. Incorporating priori information about the underlying speaker in noise or a signal-dependent model of speech signal can alleviate such limitation via decoupling the noise and speech estimators, leading to certain improvement in the perceived quality of enhanced speech possible.

Several previous attempts made steps towards incorporating a signal model or speaker model into speech enhancement framework. The authors in [1] presented an enhancement algorithm using a constrained iterative sinusoidal model as the representation of the characteristics of speech. The algorithm derived smoothed trajectories of amplitudes and frequencies and was shown to bring certain improvement in the objective speech quality measures. Other methods employed auto-regressive (AR) features as their signal model [2] or in combination with a hidden Markov model as their speaker model [3] shown to properly handle the non-stationary noise scenarios. The authors in [4] proposed to employ codebooks trained on speech and noise representing the AR-modeled spectral envelopes. Different codebooks for different noise types were trained and classified according to the observed noisy signal. The shapes of the amplitude envelopes captured in codebooks are represented by linear prediction coefficients (LPC) [5] or cepstral coefficients [6] and the appropriate gain values of the individual amplitude envelopes are obtained.

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-ICT-2011-7-288121.
presents the proposed speaker-dependent speech enhancement in sini-

2. SINUSOIDAL SIGNAL AND SPEAKER MODEL

2.1. Sinusoidal Model for Analysis/Synthesis

Let \( x(n) \) denote the time domain speech signal and its short-time
windowed segment as \( x_l(n) = x(n)w(n) \) with \( w(n) \) defined as a
window of finite support \( N \) with \( l \) as the frame index. The speech
segment \( x_l(n) \) can be well approximated as sum of sinusoids [15]:

\[
x_l(n) \approx x_l,n(n) = \sum_{d=1}^{D} A_{l,d} \cos(2\pi f_{l,d} n + \phi_{l,d}),
\]

where we define \( x_l,n(n) \) as the sinusoidal-model approximation for
\( x_l(n) \) with \( d \) as the sinusoidal index and \( D \) as the number of sinu-
soids, \( n \in [0, N-1] \) as the time index with \( N \) as the window length,
and \( A_l = \{A_{l,d}\}_{d=1}^{D}, f_l = \{f_{l,d}\}_{d=1}^{D} \) and \( \phi_l = \{\phi_{l,d}\}_{d=1}^{D} \) as the
sinusoidal amplitude, frequency and phase vectors, respectively. The
power spectrum of the sinusoidal modeled signal is given by

\[
P(\omega) \approx \sum_{d=1}^{D} A_{d}^2 W(\omega - 2\pi f_d),
\]

where \( W(\omega) \) is the magnitude response of the window \( w(n) \).

The sinusoidal features are either selected by peakpicking [16]
or by sampling the spectral envelope at harmonic multiples of an
estimated fundamental frequency [17]. In either ways, variable-
dimension sinusoidal features are inevitable due to the changes in the
fundamental frequency estimate per frame. Furthermore, due to the
fundamental frequency estimation error from noisy signal, the
harmonic multiples get erroneous, leading to wrong sinusoidal pa-
ter estimation. On the other hand, the simple peakpicking may
select wrong peaks originating from window sidelobe or noise har-
monics (see e.g. Figure 2). To avoid these problems, here we con-
sider a modified sinusoidal model which selects robust sinusoidal
features when noise is added.

The two modifications considered in the sinusoidal model are as
follows: 1) the uniform frequency division in discrete Fourier trans-
form is changed to Mel-scale by dividing the frequency range to fre-
quency bands whose center frequencies are equally distributed in Mel-
scale\(^1\), and 2) at each band, only the spectral peak with the largest
amplitude is selected. The so-obtained fixed-dimensioned sinusoidal
features is used for synthesis stage leading to a perceived qual-
ity indistinguishable from the original signal (see [18, 19] or Se-
tion 4.2 in the following for more details).

2.2. Speaker Model: Sinusoidal Codebook

To match the sinusoidal signal model, the speaker model should also
follow structure of sinusoidal coefficients as amplitude, frequency
and phase parts. As the spectral amplitude contains more perceptual
information than the phase, here we only model the sinusoidal am-
plitude and frequency for a certain speaker\(^2\).

Figure 3 shows the structure of the codebook \( C \) consisting of
\( M \) codewords each composed of an amplitude (\( C_A \)) and a frequency
(\( C_f \)) part. In the following, we denote the amplitude-frequency pair

\[
\{A^*, f^*\} = \arg \min_{A \in C_A} ||A - \hat{A}||^2,
\]

where \( ||\cdot|| \) is the Euclidean distance defined as \( ||x|| = \sqrt{\sum_{d=1}^{D} x(d)^2} \)
and \( \{A^*, f^*\} \) denote the best matching pair selected from the code-
book which minimize the Euclidean distance between the input
amplitude vector (\( A \)) and the amplitude code vector (\( A^* \)).

3. PROPOSED SINUSOIDAL-BASED SPEECH
ENHANCEMENT METHOD

The proposed method is composed of the following stages: a con-
ventional speech enhancement, a noise power estimator and a sinusoidal
dictionary of the speaker (as described in previous section) followed
by a softmask (as postfilter). The details for each step is described
in the following:

**Segmentation:** The noisy input signal \( y = x + n \) is split into short
segments using \( STFT \), where \( x \) represents the clean speech signal

\(\text{Fig. 2. Peaks chosen by sinusoidal coder (model order } D = 15 \text{) for}\)

\(\text{voiced frame: (top) clean, (bottom) noisy at SNR } = 0 \text{ (dB)}.\)

\(\text{Fig. 3. Hybrid coder structured in amplitude and frequency.}\)
and $n$ the background noise. Taking the Fourier transformation of segmented noisy speech $y(n) = y(n)w(n)$ we obtain $Y(\omega)$ with $\omega$ as frequency bin and $f$ as the frame index.

**Preprocessing:** As a preprocessing stage, Conventional speech enhancement (MMSE-LSA) [22]) is applied on each frame.

**Sinusoidal model:** Given the enhanced signal, using Eq. (3) the best matching codeword in the sinusoidal coder (C) denoted as $\theta^* = \{\hat{A}^*, \hat{f}^*\}$ is extracted as follow

$$
\theta^* = \arg \min_{\hat{A} \in C} \| \hat{A} - A \|^2_2.
$$

(4)

**Postfilter:** The postfiltering stage consists of a time-frequency softmask $G_l(\omega)$ which is driven by the decoded sinusoidal codeword $\theta^*$ and a noise estimate of the noisy input signal $\hat{\sigma}^2_{N,l}$. The found codeword $\theta^*$ is used to reconstruct a time-domain segment $\hat{x}_{l,s}$ characterized by the sinusoidal amplitude and frequency given by $\theta^*$ and the noise phase $\{\hat{\phi}_{b,d}\}_{d=1}^D$ and we have

$$
\hat{x}_{l,s}(n) = \sum_{d=1}^{D} A^*_d \cos(2\pi f^*_d n + \hat{\phi}_{b,d}).
$$

(5)

Taking the DFT of $\hat{x}_{l,s}(n)$, the power spectrum estimate for the clean speech signal can be estimated by

$$
\hat{P}_{ex}(\omega) \approx \sum_{d=1}^{D} A^*_d W(\omega - 2\pi f^*_d).
$$

(6)

Given the noise power estimate $\hat{\sigma}^2_{N,l}$ from the noise tracker, a softmask is produced as follow:

$$
G_l(\omega) = \frac{\hat{P}_{ex}(\omega)}{\hat{P}_{ex}(\omega) + \hat{\sigma}^2_{N,l}}.
$$

(7)

Applying the mask to the noisy spectrum, finally the speech spectrum is given by

$$
\hat{X}_l(\omega) = G_l(\omega) Y_l(\omega).
$$

(8)

Finally the enhanced time-domain signal denoted by $\hat{x}$ is obtained by applying ISTFT on $\hat{X}_l$ and we get

$$
\hat{x}(n) = F^{-1}\{\hat{X}_l(\omega)e^{j\hat{\phi}_b(\omega)}\},
$$

(9)

where $F^{-1}(\cdot)$ is the inverse STFT operator.

4. RESULTS

4.1. Experiment setup

As speech database we select GRID corpus [23] composed of 18 male and 16 female speakers with 1000 short sentences of the average length of 1.8 seconds. Each sentence follows a command-like structure. The speech data are downsampled from original sampling frequency 25 kHz to 8 kHz to take into account narrow band telephony speech scenarios. To simulate a noisy environment, noise recordings were used from the NOISEX-92 database [24] and were added to clean signal at specific SNR levels: $-5\text{dB}$, $0\text{dB}$, $5\text{dB}$ and $10\text{dB}$ representing very low, low, medium and high SNRs, respectively. The following noise scenarios are considered: white, babble, factory noise 1 and destroyer engine room noise.

As the frame setup, we use a Hann window of length of 32 ms and a frame shift of 8 ms. To transform the time segments into the frequency domain, a 2048-point DFT was used. As benchmark methods, we report the performance obtained by MMSE-LSA [22] and MMSE-STSA [25]. In the proposed method, it was observed that the choice of MMSE-LSA revealed better results when denoising the signal before searching the codebook. As our evaluation criteria, we choose Perceptual Evaluation of speech quality (PESQ) [26] as it was shown to have a high correlation with listening results [27]. The speaker codebooks are trained using 400 utterances per speaker disjoint from 100 utterances used to report results. The choice of Doblinger noise estimator [28] led to the best performance is fixed in our experiments. Furthermore, subjective listening tests are conducted using multiple stimuli with hidden reference and anchor (MUSHRA) test [29].

4.2. Choosing Optimal Parameters

The sinusoidal model order $D$ impacts on the perceived quality of the synthesized speech signal. If the number is too low, important spectral information is lost while if it is chosen too high, unwanted side-lobe effects of the window function are selected as sinusoidals and therefore the signal quality will degrade.

Figure 5 shows the PESQ scores versus the sinusoidal model order ($D$) for babble noise scenario. The clean sinusoidal-synthesized signal was used as our reference and to show how well the sinusoidal coder performs on clean speech signals. Furthermore, the clean sinusoidal signal is introduced, which is the time signal reconstructed from sinusoidal features without utilizing a sinusoidal codebook. Although the clean sinusoidal signal benefits from a higher model order, a choice of $D = 30$ suffices for sufficient quality as PESQ saturates for $D \geq 30$.

4.3. Spectrogram

The spectrograms are shown in Figure 6 to compare the enhanced output produced by MMSE-STSA and the proposed method, noisy signal and the clean signal. Compared to MMSE-STSA, the pro-
The proposed method has a better background noise suppression however at the expense of introducing additional artifacts caused by the quantization error and a wrong selected codeword. The PESQ scores obtained by each method are reported in the caption of the figure.

4.4. PESQ Evaluations to Assess the Perceived Quality

Figure 7 shows the average PESQ improvement results for female (left) and male (right) speakers for different noise scenarios at low (0dB), mid (5dB), and high (10dB) SNRs. We define $\Delta\text{PESQ}$ as the difference between the PESQ of the enhancement method and the PESQ of the noisy signal. For mid-SNR, PESQ improves about 0.25 (female) and 0.3 (male) in white noise scenarios. For babble noise, PESQ improvement gets to 0.2 for the female and 0.1 for the male speakers. The largest improvement obtained by the proposed method over MMSE-STSA/MMSE-LSA is achieved for very low SNR scenario, showing 0.37 improvement for both genders in white noise and 0.2 for babble noise scenario. Comparing noise-known scenario with blind one quantifies how much additional improvement is achievable when noise estimation is ideal.

Fig. 7. $\Delta\text{PESQ}$ results for (top) low SNR = 0 (dB), (middle) mid SNR = 5 (dB), and (bottom) high SNR = 10 (dB). The results are averaged over (left) female and (right) male speakers in GRID corpus.

4.5. Subjective Listening Test

To justify the objective performance evaluation results, a subjective listening test based on multiple stimuli with hidden reference and anchor (MUSHRA) instructions was conducted. MUSHRA is a double-blind multi-stimulus test method and is recommended by the International Telecommunication Union (ITU) for the evaluation of the intermediate audio quality [29]. For each scenario and every participant, a different utterance of a random speaker from the GRID corpus was used. The stimuli for the test are described in the following: i) clean reference signal (HR), ii) an anchor signal (low-pass filtered version of the clean reference signal with 2.5 kHz cutoff frequency), iii) noisy signal, iv) MMSE-STSA [25] as benchmark and v) signal processed by the proposed method. AKG K240 studio headphones were used during the listening test. Figure 8 shows the results for the MUSHRA test, averaged over all seven participants. The error bars indicate the 95% confidence interval. The proposed method performs significantly better in terms of noise reduction compared to the state-of-the-art method$^3$.

5. CONCLUSION

A single-channel speech enhancement algorithm was proposed where signal-dependent model and speaker information were incorporated. As signal model we chose sinusoidal model and as speaker model a sinusoidal coder was designed. The method pushed the limits of the conventional speech enhancement algorithm, in particular in non-stationary noise scenarios and low SNRs. Our results showed an average PESQ improvements up to 0.37 compared to state-of-the-art speech enhancement algorithms. The algorithm is real-time capable and a Matlab-based real time implementation is available to demonstrate the real-time capabilities of this method.

$^3$Wave files at: http://www.spsc.tugraz.at/iwaenc14
6. REFERENCES


