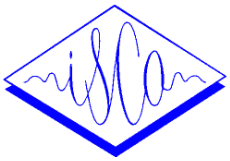

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEECH ENHANCEMENT BASED ON A CONSTRAINED SINUSOIDAL MODEL

Jesper Jensen

Center for PersonKommunikation (CPK)
Aalborg University, Denmark
jje@cpk.auc.dk

John H. L. Hansen

Robust Speech Processing Lab.
Center for Spoken Language Research
University of Colorado at Boulder, CO - USA
jhlh@cslr.colorado.edu

ABSTRACT

In this study we propose an algorithm for enhancement of speech degraded by additive broad-band noise. The algorithm represents speech using a sinusoidal model, where model parameters are estimated iteratively. In order to ensure speech-like characteristics observed in clean speech, the model parameters are restricted to satisfy certain smoothness constraints. The algorithm is evaluated using speech signals degraded by additive white Gaussian noise. Results from both objective and subjective evaluations show considerable improvement over traditional spectral subtraction and Wiener filtering based schemes. In particular, in a subjective AB preference test, where enhanced signals were encoded/decoded with the G729 speech codec, the proposed scheme was preferred over the traditional schemes in more than 5 out of 6 cases for input SNRs ranging from 5-20 dB.

1. INTRODUCTION

Often, there is a need for digital voice communications systems or automatic speech recognition systems to perform reliably in noisy environments. In many cases, these systems work well in nearly noise-free conditions, while their performance deteriorates rapidly in noisy conditions. Thus, the development of preprocessing algorithms for reducing the background degradation in speech signals is of current interest.

In the past, a number of single-microphone speech enhancement algorithms have been proposed. These include variants of spectral subtraction [2], methods based on all-pole modeling [5, 8], subspace based methods, e.g. [3, 7], and algorithms that exploit masking effects, e.g. [13]. Although successful for speech coding [9] and speech modification [12], the sinusoidal model has not received the same level of attention for speech enhancement [1, 11].

In this paper, we propose a sinusoidal model based algorithm for enhancement of speech degraded by additive broadband noise. Adopting a similar idea as that used in [5], for an all-pole model, the present algorithm exploits the notion that during the speech production process, the vocal tract transfer function varies continuously and relatively slowly with time. Furthermore, in most voiced speech regions, the fundamental frequency varies relatively slowly as well. The aim of the proposed enhancement scheme is to improve the quality of the enhanced speech signal, by exploiting this knowledge of the signals origin in the enhancement process.

The remainder of the paper is structured as follows. Sec. 2 introduces the signal model. Sec. 3 gives an overview of the enhancement algorithm, followed by a more detailed treatment of each step of the algorithm. In Sec. 4 the algorithm is evaluated using signals degraded with additive white Gaussian noise (AWGN). The evaluation is based on objective speech quality measures as well as subjective tests. Finally, Sec. 5 concludes and discusses directions for future research.

2. SIGNAL MODEL AND ASSUMPTIONS

We assume that the speech signal is corrupted by additive broadband noise, as follows,

$$\mathbf{x} = \mathbf{s} + \mathbf{n},$$

where $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{n}$ denote the noisy speech signal vector, the clean signal, and the noise component, respectively. Furthermore, it is assumed that the noise is stationary, such that the noise spectrum estimated in noise-only regions is still valid, when speech is present. The enhanced speech signal is modeled as a sum of sinusoids on a frame-by-frame basis:

$$\hat{\mathbf{s}}_m = \sum_{k=1}^{K_m} a_{k,m} \cos(\omega_{k,m} n + \phi_{k,m}), \qquad (1)$$

for $n = -N, \ldots, 0, \ldots, N$, where $\hat{\mathbf{s}}_m$ denotes the enhanced version of signal frame m, the parameters $a_{k,m}$, $\omega_{k,m}$ and $\phi_{k,m}$ denote the the k'th amplitude, angular frequency and phase, respectively, and K_m is the number of sinusoids used in frame m. Finally, since the proposed algorithm is iterative, the superscript (i) denotes iteration number, such that, e.g., $\mathbf{a}_m^{(i)}$ represents a vector containing the amplitudes of frame m at iteration i.

3. THE ENHANCEMENT ALGORITHM

3.1. Algorithm overview

It is well-known that a noise-free voiced speech signal can be modeled accurately with Eq. (1), such that the change in amplitudes and frequencies from frame to frame is usually small, provided that the parameters are updated often, say every 10 ms. That is, seen over several frames, the amplitudes and frequencies are relatively smooth functions of time. However, when estimated from a noisy signal, using e.g. peak-picking of the FFT magnitude spectrum as described in [9, p.143], the sinusoidal parameters show a much more unstructured behavior. The proposed algorithm aims at obtaining an enhanced signal, where the amplitudes and frequencies evolve smoothly with time, as would be the case in a clean voiced speech signal.

Fig. 1 outlines the proposed algorithm. First, the noisy speech signal $\mathbf{x}$ is divided into overlapping analysis frames $\mathbf{x}_m$. For each of these frames, initial sinusoidal parameter values are estimated. Amplitudes and frequencies (in voiced regions) are refined iteratively, while the estimated phase values are not modified. Finally, enhanced frames are synthesized and overlap-added in order to generate the enhanced speech signal.

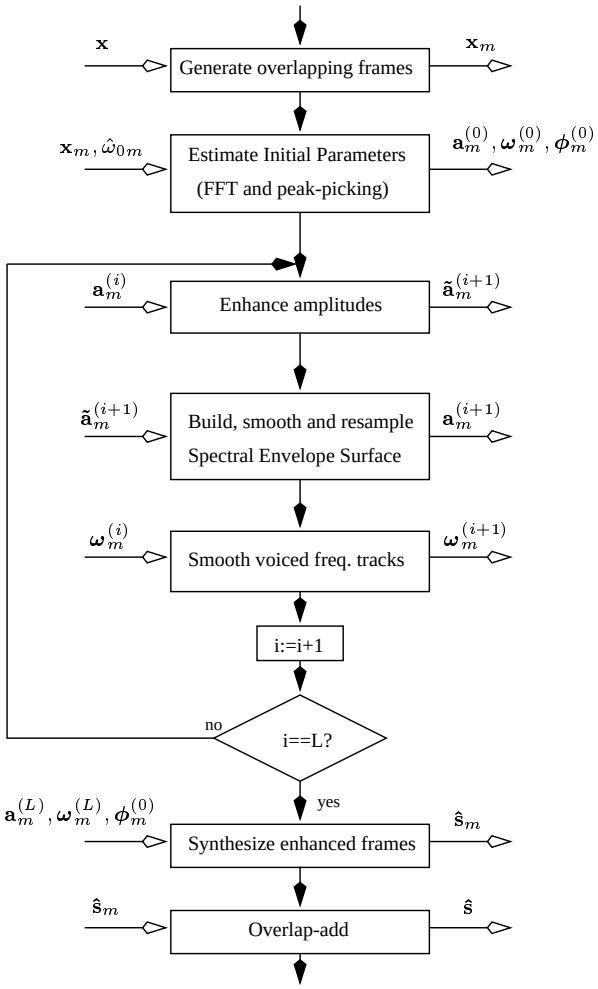

Figure 1: Block diagram of the enhancement algorithm. Black arrows: Algorithm flow, White arrows: Parameter Input/Output.

3.2. Estimation of initial parameters

Each of the noisy signal frames are Hanning windowed and transformed with a 1024 point FFT; initially all peaks of the FFT magnitude spectra are selected. These peaks represent candidate triplets $(a_{k,m}, \omega_{k,m}, \phi_{k,m})$ from which the speech signal relevant triplets are selected.

For voiced speech frames, the signal relevant peaks are due to periodicity of the speech signal, while other peaks are related to analysis window side lobes or noise. Using a rough estimate $\hat{\omega}_{0m}$ of the fundamental frequency in frame m, the frequency axis is divided into non-overlapping bands of the form $[\frac{1}{2}\hat{\omega}_{0m}; \frac{3}{2}\hat{\omega}_{0m}), [\frac{3}{2}\hat{\omega}_{0m}; \frac{5}{2}\hat{\omega}_{0m}), \ldots$, and a peak from each band is selected. Since the noise is assumed stationary, a SNR for each frequency band can be estimated. In bands with SNR above a prespecified value, T_{SNR}, the highest peak is selected, while in all other bands, the peak closest to the center of the frequency band is selected. Using this peak-picking strategy in low SNR bands generally works better than selecting the highest peak in all bands.

For unvoiced frames it is more difficult to decide which peaks to retain. However, typically, some voiced frames will be misclassified as unvoiced when noise is present. In this case, the side lobe peaks between the harmonic peaks should be discarded in order to avoid audible artifacts in the enhanced signal. A simple procedure to determine if a peak in an unvoiced frame is a side lobe peak is to calculate the slopes of the lines between the peak and its two neighboring peaks. If the neighboring peaks have higher amplitude and the slopes are steeper than a pre-specified value T_{unv}, the peak is a side lobe peak and should be discarded.

Using this peak-picking procedure for noise-free speech signals results in modeled signals that are almost indistinguishable from the originals.

3.3. Enhancement and smoothing of amplitudes

The procedure for enhancement of sinusoidal amplitudes consists of two steps. The first step aims at reducing the noise, while the second step ensures that amplitudes evolve smoothly with time.

Noise is reduced in an iterative Wiener filtering scheme, where enhanced amplitudes are estimated using a weighted average between amplitude values from the previous iteration and their Wiener filtered counterparts:

$$\tilde{a}_{k,m}^{(i+1)} = W_w H_{k,m} a_{k,m}^{(i)} + (1 - W_w) a_{k,m}^{(i)} \quad ,$$

where $H_{k,m}$ is the value of the Wiener filter at frequency $\omega_{k,m}$,

$$H_{k,m} = \frac{(a_{k,m}^{(i)})^2}{(a_{k,m}^{(i)})^2 + N_k^2},$$

and N_k denotes the noise amplitude spectrum at frequency $\omega_{k,m}$. The weight factor W_w controls the amount of noise reduction at each iteration.

Using the enhanced amplitude values, an approximation of the vocal tract spectral envelope for frame m is obtained by linear interpolation between the spectral peaks $\left(\omega_{k,m}^{(i)}, \log(\tilde{a}_{k,m}^{(i+1)})\right), k = 1, \ldots, K_m$, in the log-amplitude domain, see e.g. [9, p. 143]. The spectral envelope approximation for frame m is denoted $S_{j,m}^{(i)}, j = 1, \ldots, J$. Such spectral envelope approximations for consecutive frames constitute the so-called Spectral Envelope Surface (SES), which is actually a grid of $S_{j,m}^{(i)}$ points. Since the vocal tract transfer function presumably varies continuously and relatively slowly with time, and the SES is an approximation of this, the SES should be a smooth surface.

For this reason, a smoothing procedure is applied to the estimated SES. The SES is smoothed by calculating a weighted average between each of the points on the SES and the eight neighboring points. The weight parameter in this averaging procedure is SNR-dependent, and selected such that SES points at high SNR are not greatly modified, while SES points at low SNR can change more during the smoothing process. After smoothing of the SES, enhanced and smoothed amplitude values $\mathbf{a}_{k,m}^{(i+1)}$ are obtained by resampling the smoothed SES at points corresponding to the frequencies $\omega_{k,m}$.

3.4. Smoothing of frequencies

In noisy conditions, only a rough estimate of the fundamental frequency can be expected to be available. This, combined with the peak-picking procedure described above, results in rough frequency tracks in voiced regions, particularly at higher frequencies. However, noise-free voiced speech can be represented with the sinusoidal model, such that the frequency tracks evolve smoothly with time. For this reason, a smoothing procedure is applied to the sinusoidal frequencies in voiced regions. Each frequency in the current frame is linked to a frequency in the previous and the following frame, and a smoothed frequency value is calculated as a weighted average. This smoothed frequency value is used, unless the relative frequency change from frame to frame exceeds a threshold T_f (informal evaluations showed that $T_f = 20\%$ performed well). In this case, a pitch halving/doubling error probably has occurred, and the original unsmoothed frequency value is kept.

3.5. Termination criterion

An open question is when to terminate the iterative process. This is a trade-off between noise reduction and speech distortion. If the iterations are too few, the enhanced signal is noisier than necessary, and if too many iterations are performed, portions of the speech signal could have artifacts introduced.

In order to find a scheme for determining when to terminate the iterations, enhanced speech signals were synthesized after each iteration of the algorithm. The objective speech quality of each of the enhanced signals was estimated by calculating the symmetric log-likelihood ratio (llr) measure defined in [10, pp.49], which has been shown to correlate fairly well with subjective quality [10].

Initially, the optimum iteration number (in terms of llr) varied both with input SNR and speakers. However, it turned out, that with a proper selection of the weight factors used in the enhancement and smoothing procedures described above, optimum enhanced signals could be achieved after $L = 6$ iterations, almost independent of speakers and input noise level.

3.6. Signal re-synthesis

The iterative scheme described above results in enhanced/smoothed amplitudes and frequencies. Enhanced signal frames are generated by inserting these improved parameters and the original noisy phase values into Eq. (1). Next, the enhanced speech signal is synthesized by overlap-adding the enhanced frames using a triangular window as the synthesis window.

4. ALGORITHM EVALUATION

Test sentences, sampled at 8 kHz, were selected randomly from the TIMIT database, and degraded with AWGN at SNR levels of 20, 15, 10 and 5 dB. The SNR is defined as:

$$\text{SNR [dB]} = 10 \log_{10}(\mathbf{s}^T \mathbf{s} / \mathbf{n}^T \mathbf{n}),$$

where $\mathbf{s}$ represents an entire sentence, $\mathbf{n}$ is the corresponding noise sequence, and T denotes vector transposition.

For the proposed enhancement scheme, voicing decisions were made based on frame energy and zero-crossing rate. In voiced frames, fundamental frequency estimates were obtained using an improved correlation based pitch estimator. The enhancement algorithm used analysis frames of a length of 200 samples (25 ms) and new frames were selected every 80 samples (10 ms). The triangular window for overlap-add synthesis had a length of 160 samples (20 ms).

4.1. Objective Evaluation

Four test sentences, two female and two male, were enhanced with the proposed algorithm, where a fixed number of $L = 6$ iterations was used. For comparison, the test sentences were enhanced with the spectral subtraction method in [2] using magnitude averaging, and with the unconstrained Wiener filtering approach in [8] as well. In the latter scheme, up to 10 iterations were performed for each signal, and the optimum enhanced signals in terms of llr were selected.

For objective quality assessment of the test sentences, the average symmetric llr value was calculated from frames of length 240 samples taken with an overlap of 75% throughout each sentence.

Fig. 2 compares the objective speech quality of the enhanced signals obtained from the different schemes. Fig. 2a, 2b, and 2c show average llr values from unvoiced regions, voiced regions, and non-silence regions (i.e. unvoiced, voiced and transitional regions), respectively.

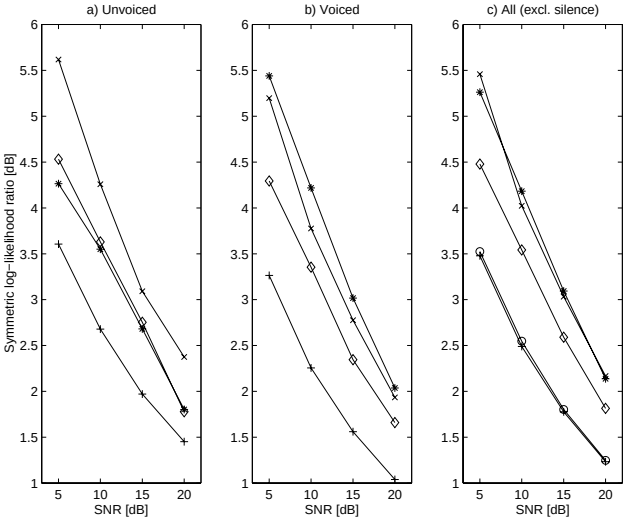

Figure 2: Enhancement performance in terms of llr. a) Unvoiced segments, b) Voiced segments, c) All segments (excluding silence).
Noisy signals: $*$. Spectral Subtraction: x. Unconstrained Wiener: $\diamond$. Proposed method with F0 and voicing from clean signals: +. Proposed method with F0 and voicing from noisy signals: $\circ$.

Fig. 2a shows that in unvoiced segments, spectral subtraction and unconstrained Wiener filtering perform poorly, while some improvement can be observed with the proposed method. In voiced segments, Fig. 2b, all methods improve speech quality, but the proposed scheme reaches lower llr-values than the traditional methods. Fig. 2c shows that, generally, better performance can be obtained with the proposed enhancement scheme. Furthermore, using F0 and voicing estimates based on noisy signals, which would be the case in any practical situation, does not seem to degrade over-all performance significantly. In particular, the average llr when using noisy F0 estimates is only 1% higher than with clean signals.

4.2. Subjective Evaluation

In general, signals enhanced with the proposed scheme generally have high subjective quality in voiced regions. Here, the residual noise is typically a low-level, high-frequency 'ringing', somewhat similar to 'musical noise'. This residual noise can partly be contributed to the noisy phases used in the re-synthesis process.

In unvoiced regions, the enhanced signal seems to have lower subjective quality; in particular, stops tend to sound slightly 'muffled'. This is due to several factors. First, the local SNR in unvoiced regions is typically much lower than in voiced regions. Secondly, the sinusoidal model is not well-suited for representing signal segments, such as stops, where the amplitude level changes rapidly with time, see [6].

In order to study the influence of inaccurate voicing and F0 information, noisy signals were enhanced using F0 and voicing estimates from clean and noisy signals, respectively. Informal listening confirms the objective test results from Fig. 2c, that the algorithm is not particularly sensitive to accurate F0 and voicing information, i.e., a rough F0-contour is adequate for near-optimum performance. However, the algorithm seems sensitive to bursts of F0 halving/doubling errors; sometimes, such bursts cause audible artifacts in the enhanced speech signal. This does not occur often, with between 10-15 frames containing pitch errors for the +5600 frames employed in the enhancement evaluation. Also, such errors almost always occured for high pitch female speakers. Further effort in detection of pitch tracking errors would help eliminate this issue.

4.3. AB-comparison test with G729-coded speech

Finally, the proposed scheme was evaluated as a front-end for a speech codec in an informal AB-preference test. Three degraded test signals, different from the signals used in the previous section, were enhanced with the proposed scheme using F0 and voicing estimates from the noisy signals; subsequently, the enhanced signals were encoded/decoded with the G729 8 kbit/s CS-ACELP speech codec [4]. Ten subjects were asked to compare these signals to encoded/decoded versions of the original noisy signals, signals enhanced with spectral subtraction, and signals enhanced with unconstrained Wiener filtering. The preference towards the proposed scheme is shown in Table 1.

AB Preference Test			
SNR [dB]	Noisy Signal	Spec.Sub.	Unconstr. Wiener
20	29/30	30/30	30/30
15	30/30	30/30	30/30
10	30/30	30/30	30/30
5	29/30	30/30	26/30

Table 1: Preference for the proposed scheme vs. G729-encoded/decoded versions of the original noisy signal, spectral subtraction, and unconstrained Wiener filtering.

From this table it is clear, that the proposed enhancement scheme is generally preferred over the noisy, original signals, spectral subtraction and unconstrained Wiener filtering.

5. CONCLUSIONS AND FUTURE WORK

An iterative sinusoidal model-based scheme has been proposed for enhancement of speech degraded by additive broad-band noise. Smoothness constraints were imposed on sinusoidal amplitudes and frequencies (in voiced regions), in order to ensure a parameter behavior similar to that of clean speech. Objective and subjective improvements were observed compared to spectral subtraction and unconstrained Wiener filtering.

Further improvement of performance in unvoiced speech regions is a topic for future research. This could, e.g., be done by introducing another signal model in these regions. Furthermore, alternatives to the computationally intensive parameter smoothing scheme are currently being studied.

6. REFERENCES

[1] D. V. Anderson and M. A. Clements. Audio Signal Noise Reduction Using Multi-Resolution Sinusoidal Modeling. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 805–808, 1999.

[2] S. F. Boll. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

[3] Y. Ephraim and H. L. Van Trees. A Signal Subspace Approach for Speech Enhancement. *IEEE Trans. Speech, Audio Processing*, 3(4):251–266, 1995.

[4] ITU-T Rec. G.729. Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Predictive (CS-ACELP) Coding.

[5] J. H. L. Hansen and M. A. Clements. Constrained Iterative Speech Enhancement with Application to Speech Recognition. *IEEE Trans. Signal Processing*, 39(4):795–805, April 1991.

[6] J. Jensen, S. H. Jensen, and E. Hansen. Exponential Sinusoidal Modeling of Transitional Speech Segments. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 473–476, 1999.

[7] S. H. Jensen et al. Reduction of Broad-Band Noise in Speech by Truncated QSVD. *IEEE Trans. Speech, Audio Processing*, 3(6):439–448, 1995.

[8] J. S. Lim and A. V. Oppenheim. All-Pole Modeling of Degraded Speech. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-26(3):197–210, 1978.

[9] R. J. McAulay and T. F. Quatieri. Sinusoidal Coding. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 4. Elsevier Science B. V., 1995.

[10] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[11] T. F. Quatieri and R. J. McAulay. Noise Reduction using a Soft-Decision Sine-Wave Vector Quantizer. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 821–824, 1990.

[12] T. F. Quatieri and R. J. McAulay. Shape Invariant Time-Scale and Pitch Modification of Speech. *IEEE Trans. Signal Processing*, 40(3):497–510, 1992.

[13] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis. Speech Enhancement based on Audible Noise Suppression. *IEEE Trans. Speech, Audio Processing*, 5(7):497–513, 1997.