ADAPTIVE TIME SEGMENTATION OF NOISY SPEECH FOR IMPROVED SPEECH ENHANCEMENT

Richard C. Hendriks, Richard Heusdens and Jesper Jensen

Dept. of Mediamatics
Delft University of Technology
2628 CD Delft, The Netherlands
email: {R.C.Hendriks, R.Heusdens, J.Jensen}@EWI.TUDelft.nl

ABSTRACT

Enhancement algorithms are widely used to overcome the degradation of noisy speech signals. Most enhancement algorithms require an estimate of the noise and noisy speech power spectra in order to compute the gain function used for the noise suppression. The variance of these power spectral estimates degrades the quality of the enhanced signal and smoothing techniques are therefore often used to decrease the variance. In this paper we present a method to determine the noisy speech power spectrum based on an adaptive time segmentation. More specifically, the proposed algorithm determines for each noisy frame which of the surrounding frames should contribute to the corresponding noisy power spectral estimate. Objective and subjective experiments show that an adaptive time segmentation leads to significant performance improvements, particularly in transitional speech regions.

1. INTRODUCTION

The need for single-channel enhancement of speech signals degraded by noise arises frequently in mobile communication applications. Within single-channel speech enhancement the noise is often assumed additive, i.e. $y = x+n$, with y the noisy speech signal, x the clean speech signal and n the noise realization. Recently, the class of frequency domain enhancement methods have received significant interest partly due to their relatively good performance and low computational complexity. These methods transform the noisy speech signal frame by frame to the spectral domain, e.g. using a Discrete Fourier Transform (DFT). Here, complex-valued DFT coefficients of the clean signal are estimated by applying a gain function (e.g. the Wiener [1] or LSA gain [2]) to the noisy DFT coefficients. Subsequently enhanced time domain frames are generated using the inverse DFT and the enhanced waveform is constructed by overlap-adding the enhanced frames.

Gain functions are typically computed from two quantities, namely an estimate of the noise power spectrum and of the noisy speech power spectrum. While the problem of estimating and tracking the noise power spectrum in speech presence has received significant interest recently [3], methods for accurate estimation of the noisy speech power spectrum appear to have been less explored. Classical methods for estimating the noisy speech power spectrum include the periodogram, computed as $\frac{1}{N}|Y(f)|^2$, where

The research is supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

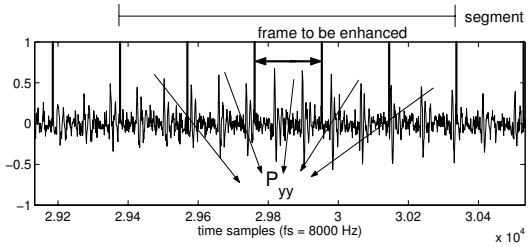

Fig. 1. *Noisy speech signal with frame to be enhanced. In this example a segment consists of 5 consecutive frames.*

$Y(f)$ is the Fourier transform of the noisy speech sequence $y(n)$. However, this estimator suffers from a variance of $var[\hat{P}_{yy}(f)] \propto P^2_{yy}(f)$ [4] with P_{yy} the power spectrum of the noisy speech and $\hat{P}_{yy}$ its estimate. To decrease the variance of the estimated noisy speech power spectrum, smoothing methods like Bartlett's method can be used [4]. These methods offer an estimated (smoothed) power spectrum by averaging periodograms of, say K consecutive frames, hereby decreasing the variance of the power spectrum estimate by a factor K [4]. However, the decrease in variance comes with a side effect: the frequency resolution is decreased as well.

In order to apply smoothing based methods for estimation of P_{yy}, smoothing must be performed within stationary segments. Each segment may consist of a number of frames including the frame to be enhanced, as shown in Fig. 1. In Boll's work on spectral subtraction [5], smoothing was performed across segments consisting of 3 frames located symmetrically around the frame to be enhanced. However, ideally, segments should vary with speech sounds: some vowel sounds may be considered stationary up to 40-50 ms, while stop consonants may be stationary for less than 5 ms [6]. In existing enhancement systems the length of segments is typically fixed and reflects the *average* stationary duration of speech sounds, typically 20 ms. Using a fixed segment size has two potential drawbacks. First, in signal regions which can be considered stationary for longer time than the segment used, the variance of the spectral estimator is unnecessarily large. Secondly, if stationarity of the speech sound is shorter than this fixed segment size, smoothing is applied across stationarity boundaries resulting in blurring of transients and of rapidly-varying speech components [7], leading to a degradation of the speech intelligibility.

In [7] a method was presented to overcome the two described problems using an adaptive exponential smoother. Here the amount of smoothing was adapted to the underlying speech process by us-

ing a measure reflecting the degree of stationarity based on spectral derivatives. In this paper, we propose a different approach to overcome the two above mentioned problems, namely by using an adaptive time segmentation for speech enhancement. To be more specific, the proposed method determines which noisy speech data should contribute in the estimation of the noisy speech power spectrum for a given frame. The proposed algorithm is very general. It can work as a front-end for most existing speech enhancement systems and is independent of the particular suppression rule (e.g. Wiener, LSA, etc.) that is used in the enhancement algorithm.

2. ADAPTIVE TIME SEGMENTATION

The segmentation algorithm we propose here is based on a probabilistic framework, where segments are formed based on the outcome of a hypothesis test. We test the hypotheses whether two consecutive sequences of time-samples should be merged to form one segment or not. Here we regard sequences of time samples as an outcome of random processes and search for sequences that are stationary to a certain degree. In particular, we will use a test statistic based on a necessary condition for stationarity, namely that zero-lag correlation coefficients of the random process must remain invariant over time. This means that $R[0] = E\left\{|y(n)|^2\right\}$, with $R[0]$ the energy or correlation coefficient with lag 0 should be constant over time. Let s_1 and s_2 be union of speech samples from frames $i = n, ..., n + n_0 - 1$ and $i = n + n_0, ..., n + N - 1$ respectively. The two hypotheses then are:

H_0 : $R[0]$ in s_1 and $R[0]$ in s_2 are drawn from the same distribution

H_1 : $R[0]$ in s_1 and $R[0]$ in s_2 are drawn from different distributions.

The decision between the two hypotheses is made based on the following likelihood ratio test (LRT) [8],

$$\text{Reject } H_0 \text{ if} \qquad \frac{p(R[0]|H_1)}{p(R[0]|H_0)} > \gamma, \tag{1}$$

with γ a decision threshold, and $p(R[0]|H_0)$ and $p(R[0]|H_1)$ the pdf of $R[0]$ under hypothesis H_0 and H_1 respectively. For a given choice of γ, (1) is known as the Neyman-Pearson test, which maximizes the detection probability $P(H_1|H_1)$ for a given false alarm probability $P(H_1|H_0)$. In order to apply Eq. (1), pdfs $p(R[0]|H_1)$ and $p(R[0]|H_0)$ must be determined. We will argue that under certain assumptions the type of those pdfs is Gaussian and use the standard procedure of the Generalized LRT [8] and substitute unknown pdf parameters with their maximum likelihood estimates.

2.1. Distribution of $R[0]$

The central limit theorem states that the normalized sum of a large number of mutually independent random variables $X_1, ..., X_N$ with zero means and finite variances $\sigma_1^2, ..., \sigma_N^2$ tends to the normal probability distribution provided that the individual variances σ_n^2, $n = 1, ..., N$ are small compared to $\sum_{n=1}^{N} \sigma_n^2$ [9]. To determine the distribution type of $R[0]$ we assume that the time samples are independent random variables (as is commonly done in speech enhancement [2]). Because $R[0]$ can be estimated as

$$\hat{R}[0] = \frac{1}{N} \sum_{k=1}^{N} y^2(n),$$

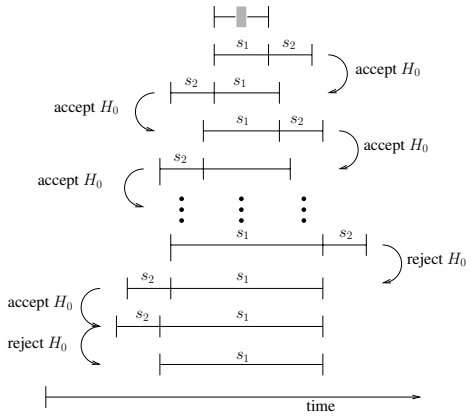

Fig. 2. *Segmentation algorithm based on hypothesis.*

$R[0]$ is a sum of independent random variables and it follows then that $R[0]$ is Gaussian distributed.

2.2. Segmentation Procedure

Knowing the distribution of $R[0]$, we are now in a position to compute the likelihood ratio given sequences s_1 and s_2,

$$\frac{p(R[0]|s_1, s_2, H_1)}{p(R[0]|s_1, s_2, H_0)} = \frac{p(R[0]|s_1, H_1)p(R[0]|s_2, H_1)}{p(R[0]|s_1, s_2, H_0)}.$$

In principle, to find for a given frame a corresponding segment, we should perform an exhaustive search over all possible segments. To avoid this computationally demanding full-search approach, we propose instead a computationally simpler algorithm which simulation experiments have shown to lead to the same performance as the full search algorithm. In Fig. 2 this simplified algorithm is described. Start with a minimum segment s_1, which is assumed to be stationary and contains the frame under consideration (shaded area in Fig. 2). Then extend this minimum segment with one frame at a time in an iterative process. Whether the segment should be extended with a neighboring frame is decided using the hypothesis test over sequence s_1 and a neighboring sequence s_2. We continue this process until on both sides of s_1 H_0 is rejected. The final sequence s_1 is considered as the stationary segment that can be used for smoothing of the noisy speech power spectrum.

This segmentation algorithm can be generalized by dividing the frequency range in sub-bands and determine a segmentation for each band independently. However, in this case less information is present per band to do a maximum likelihood estimation of the mean and variance. This, in turn, means that the variance of these estimates will be larger than in the full-band case. We expect that increasing the number of bands may be beneficial for a small number of bands, but for larger number of bands the advantage of having many bands may be overshadowed by the increased variance of the spectral estimate in each band.

Fig. 3 shows a block scheme of the proposed segmentation algorithm in combination with an enhancement algorithm. First a noisy frame y_i is divided into L frequency bands. Then for each frequency band a segmentation is determined that is used to estimate the noisy power spectrum in that band. Then a gain function (e.g. Wiener, LSA, etc.) is calculated based on this noisy power spectral estimate, and an estimate of the noise power spectrum, which we assume is available. Finally, the FFT of the noisy speech

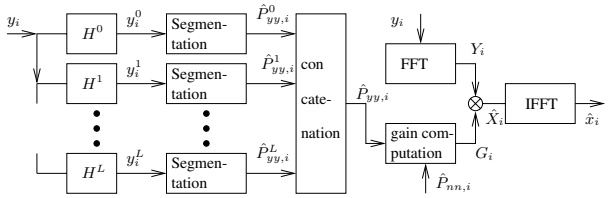

Fig. 3. *Block diagram of flexible segmented speech enhancement system.*

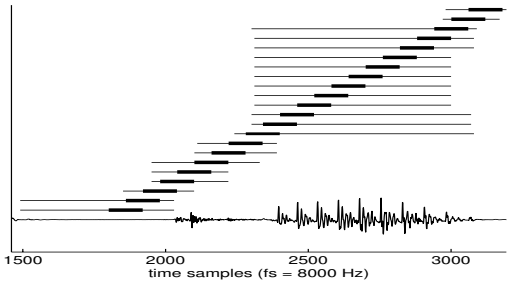

Fig. 4. *Example Segmentation. Thick horizontal lines: Duration of frames. Thin horizontal lines: Corresponding segments.*

frame y_i is scaled by the gain function, and the enhanced frame $\hat{x}_i$ is computed using an IFFT.

In Fig. 4 we show the result of the above described hypothesis based segmentation algorithm applied to a speech signal degraded by white noise at an SNR of 15 dB. In the figure the original clean speech signal is shown together with the resulting segmentation. The thick lines mark the frames in which the signal is divided for enhancement. The thin lines represent for each frame the corresponding segment that is found by the hypothesis based algorithm. In this example we used a full-band version of the above described algorithm. The speech signal under consideration consists of four parts. An initial silence part, a transient, some ringing after the transient and a voiced part. We see that frames in the silence and voiced part have long segments associated which cover respectively the whole silence and voiced part. Frames in the transient part have rather short segments. This prevents smearing of the transient. Further the beginning of the voiced part is resolved, preventing it from smearing into the ringing of the transient.

3. IDEAL SEGMENTATION

In order to obtain a bound of the performance of our segmentation algorithm presented in Section 2 we consider in this section an idealized situation, where optimal segments are found using knowledge of the clean signal. Clearly, in practical situations such an approach is not possible. The ideal segmentation is found by

$$\min_{s \epsilon S} E[D(x, \hat{x}(s))], \tag{2}$$

where s is a segmentation from the set S of all allowed segmentations, x the clean speech signal, $\hat{x}$ the estimated clean speech signal, D a distance measure between the clean speech signal x and the estimated clean speech signal and E the expectation operator. The expectation operator is used to eliminate the influence of the noise realization on the distortion measure. By doing so, the distortion is not optimized for a particular realization of the noise but for its statistical properties. We assume that distortions across frames are additive and independent. We can then write (2) as

$$\sum_i \min_{s_i \epsilon S_i} E[D(x_i, \hat{x}_i(s_i))], \tag{3}$$

where i is the frame index, x_i is the clean speech frame, $\hat{x}_i$ the estimated clean speech frame and s_i a certain segment from the set of all allowed segments for frame with index i. The purpose of (3) is to find for each frame a corresponding segment such that D is minimized. The distortion measure we minimize here is the l_2 difference between the clean and the estimated signal;

$$D = \|X_i - G_i Y_i\|^2, \tag{4}$$

where $X_i \in \mathbb{C}^M$ and $Y_i \in \mathbb{C}^M$ are in the Fourier domain, $G_i \in \mathbb{R}^{M \times M}$ is a linear filter matrix, and M the FFT order. Like the hypothesis based segmentation described in Section 2, the ideal segmentation can also be generalized by dividing the frequency range in sub-bands. In contrast to the hypothesis based segmentation, increasing the number of bands for the ideal case will result in a better segmentations always.

4. SUBJECTIVE AND OBJECTIVE RESULTS

We evaluate the presented segmentation algorithms by means of objective and subjective experiments. We use the segmentation algorithms as front-ends for Wiener filter based enhancement algorithms with a gain function $G = (R_{yy} - R_{nn})R_{yy}^{-1}$, with R_{yy} and R_{nn} the autocorrelation function of the noisy speech and noise respectively. In Fig. 5 the impact of our flexible hypothesis based segmentation algorithm is demonstrated on a female speech signal and compared with a fixed segmentation. The duration of the segments for the fixed segmentation was 53.8 ms and the threshold for the hypothesis based segmentation was $\gamma = 10^{6.9}$, both based on an optimal segmental SNR of simulation experiments over 30 different speech sentences. Segmental SNR is defined as $\frac{1}{N}\sum_{i=0}^{N-1} 10\log_{10}\frac{\|x_i\|^2}{\|x_i - \hat{x}_i\|^2}$, with N the number of frames [6]. In Fig. 5 we show the clean speech signal together with the SNR per frame after enhancement of a noisy speech signal for both the fixed segmentation and the hypothesis based algorithm with four sub-bands. Here clean speech was degraded by white noise at an SNR of 15 dB. Especially at the locations where the speech signal changes abruptly, the presented hypothesis based segmentation improves performance in terms of output SNR. The local improvements of 10 dB around the beginnings and endings of speech sounds are due to less smearing of the speech sound.

As a second objective evaluation we compared fixed segmentation, ideal flexible segmentation and the hypothesis based segmentation algorithm in terms of segmental SNR. The fragments were sampled at 8 kHz. Frame sizes of 120 samples with 50 percent overlap were used. The results are averaged over 6 different speakers, 3 male and 3 female speakers and are shown in Fig. 6a and 6b for input SNRs of 5 and 15 dB, respectively. Both figures show the segmental SNR versus the average segment length for the fixed segmentation, the hypothesis based segmentation with 4 equal width sub-bands and the ideal segmentation with 1, 2, 4, 8 and 16 equal width sub-bands. The segment length for the fixed segmentation and the threshold γ for the hypothesis based segmentation were again based on optimal segmental SNR of simulation experiments over 30 different speech sentences. For the hypothesis based segmentation we used 4 sub-bands, because that led to an

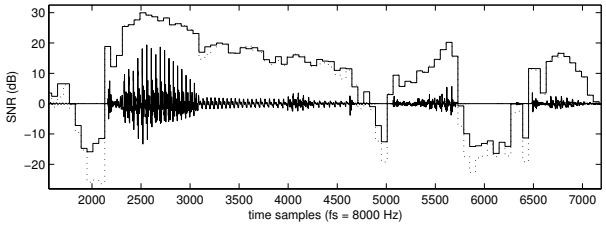

Fig. 5. *SNR per frame after using the fixed segmentation (dotted) with segment length of 53.8 ms and the hypothesis based segmentation with four sub-bands (solid) with threshold $\gamma = 10^{6.9}$. Input SNR was 15 dB.*

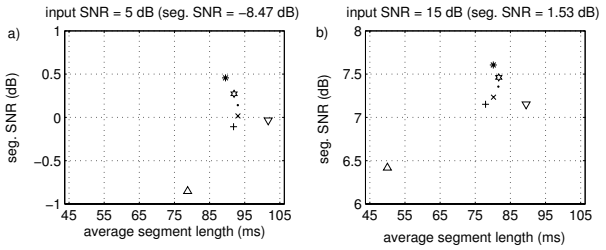

Fig. 6. *Comparison between fixed segmentation ($\triangle$), ideal segmentation with 1 (+), 2 (x), 4 ($\cdot$), 8 ($\star$) and 16 ($*$) sub-bands and hypothesis based segmentation with 4 sub-bands ($\bigtriangledown$). a): Input SNR of 5 dB. b): Input SNR of 15 dB.*

optimal segmental SNR. Comparing Fig. 6a and 6b it can be seen that for lower input SNR, all methods have longer segment lengths. The segmental SNR of the hypothesis based segmentation with 4 sub-bands is 0.82 dB and 0.73 dB higher than the segmental SNR of the fixed segmentation for respectively an input SNR of 5 and 15 dB. Further it can be seen that hypothesis based segmentation with 4 sub-bands has approximately the same segmental SNR as the full-band version of the ideal segmentation. From Fig. 6 it is clear that the segmental SNR for the ideal segmentation increases with the number of sub-bands that is used.

In the experiments reported so far we used the same value of the threshold γ for each frame and for each frequency band. We would, however, expect a performance gain if we allow different γ-values for different frames or frequency bands dependent on the SNR. However, by experiments over 30 different speech signals, it was observed that the optimal γ is fairly insensitive to the SNR.

For subjective evaluation an OAB listening test was performed with nine participants, the authors not included. Here, O is the original signal and A and B are two enhanced signals. We implemented a Wiener filter combined with a fixed segmentation and a Wiener filter combined with the hypothesis based segmentation algorithm with 4 sub-bands. Six speech signals were used, 3 male and 3 female speakers, all degraded by white noise at an SNR of 5 and 15 dB. We presented the listeners first the original signal followed by two versions enhanced with a fixed or hypothesis based segmentation. Each series was repeated 4 times with the enhanced versions played in random order. The results for all signals are shown in Table 1. The percentages represent the relative preference of the hypothesis based segmentation, with the standard deviation between brackets. A student's t-test confirmed that at an SNR of 5 dB the preference of the proposed algorithm was statistically significant, except for signal number 5. At 15 dB SNR the difference was significant for all signals except number 4 and 5.

signal no.	input SNR 5 dB	sign. 5%	input SNR 15 dB	sign. 5%
1	86.1% (28.6)	yes	88.9% (18.2)	yes
2	83.3% (17.7)	yes	80.6% (32.5)	yes
3	75.0% (21.7)	yes	77.8% (29.2)	yes
4	77.8% (23.2)	yes	63.9% (28.6)	no
5	52.8% (42.3)	no	58.3% (46.8)	no
6	88.9% (18.2)	yes	75.0% (21.7)	yes

Table 1. *Listening test results for 5 and 15 db input SNR.*

5. CONCLUSIONS

We presented an adaptive time segmentation for speech enhancement to improve the estimation of the noisy speech power spectrum. The segmentation algorithm only needs knowledge of the noisy speech signal to determine for each frame which segment of data should be used to estimate the noisy speech power spectrum. The segments are formed based on the outcome of a sequence of hypothesis tests. Objective experiments showed that usage of the adaptive time segmentation leads to a better quality in terms of SNR and that the performance of hypothesis based segmentation is close to that obtained with ideal segmentation. Furthermore, subjective listening tests showed that in terms of perceptual quality the adaptive segmentation algorithm is preferred over the usage of a fixed segmentation.

6. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[4] J. G. Proakis, C. M. Rader F. Ling C. L. Nikias, M. Moonen, and I. K. Proudler, *Algorithms for Statistical Signal Processing*, Prentice Hall, Upper Saddle River, NJ, 2002.

[5] S. F. Boll, "Suppression of acoustic noise in speec using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[6] J.R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, Piscataway, NJ, 2000.

[7] T. F. Quatieri and R. B. Dunn, "Speech enhancement based on auditory spectral change," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, May 2002, vol. 1, pp. 257–260.

[8] H. L. van Trees, *Detection, Estimation and Modulation Theory*, vol. 1, John Wiley and Sons, 1968.

[9] H. Stark and J.W. Woods, *Probability, random processes, and estimation theory for engineers*, Prentice Hall, Englewood Cliffs, NJ, 1986.