

Nonlinear Enhancement of Onset for Robust Speech Recognition

Chanwoo Kim¹ and Richard M. Stern²

Language Technologies Institute¹
and Department of Electrical and Computer Engineering^{1,2}
Carnegie Mellon University, Pittsburgh, PA 15213 USA

chanwook@cs.cmu.edu, rms@cs.cmu.edu

Abstract

In this paper we present a novel algorithm called Suppression of Slowly-varying components and the Falling edge of the power envelope (SSF) to enhance spectral features for robust speech recognition, especially in reverberant environments. This algorithm is motivated by the precedence effect and by the modulation frequency characteristics of the human auditory system. We describe two slightly different types of processing that differ in whether or not the falling edges of power trajectories are suppressed using a lowpassed power envelope signal. The SSF algorithms can be implemented for online processing. Speech recognition results show that this algorithm provides especially good robustness in reverberant environments.¹

Index Terms: Robust speech recognition, speech enhancement, precedence effect, modulation frequency

1. Introduction

Despite continued efforts by a large number of researchers, enhancing the noise robustness of automatic speech recognition systems still remains a very challenging problem. For stationary noise such as white or pink noise, algorithms such as histogram normalization (e.g. [1]) or vector Taylor series (VTS) [2] have been shown to be effective. Nevertheless, similar improvement has not been observed in more realistic environments such as background music [3]. In these more difficult environments, it has been frequently observed that algorithms motivated by human auditory processing (e.g. [4]) or missing feature algorithms (e.g. [5]) are more promising.

It has long been believed that modulation frequency plays an important role in human hearing. For example, it is observed that the human auditory system is more sensitive to modulation frequencies less than 20 Hz (e.g. [6, 7]). On the other hand, very slowly changing components (e.g. less than 5 Hz) are usually related to noise sources (e.g. [8, 9, 10]). Based on these observations, researchers have tried to utilize modulation frequency information to enhance the speech recognition performance in noisy environments. Typical approaches use highpass or band-pass filtering in either the spectral, log-spectral, or cepstral domains (e.g. [11]). Hirsch *et al.* [12] investigated the effects of highpass filtering of spectral envelopes of each frequency sub-band. Hirsch conducted highpass filtering in the log spectral domain using the transfer function:

$$H(z) = \frac{1 - z^{-1}}{1 - 0.7z^{-1}} \quad (1)$$

¹This work was supported by the National Science Foundation (Grant IIS-10916918).

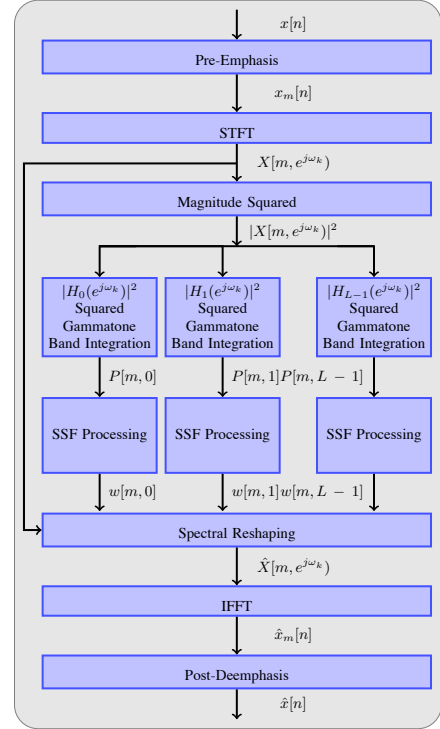


Figure 1: Block diagram of the SSF processing system.

For robust speech recognition, the other common difficulty is reverberation. Many hearing scientists believe that human speech perception in reverberation is enabled by the “precedence effect”, which refers to an emphasis that appears to be given to the first-arriving wave-front of a complex signal in sound localization and possibly speech perception (e.g. [13]). To detect the first wave-front, we can either measure the envelope of the signal or the energy in the frame (e.g. [14, 15]).

In this paper we introduce an approach which we refer to as SSF processing, representing *Suppression of Slowly-varying components and the Falling edge* of the power envelope. This processing mimics aspects of both the precedence effect and modulation spectrum analysis. SSF processing operates on frequency-weighted power coefficients as they evolve over time, as described below. The DC-bias term is first removed in each frequency band by subtracting an exponentially-weighted moving average. When the instantaneous power in a given frequency channel is smaller than this average, the power is suppressed, either by scaling by a small constant, or by replacement by the scaled moving average. The first approach re-

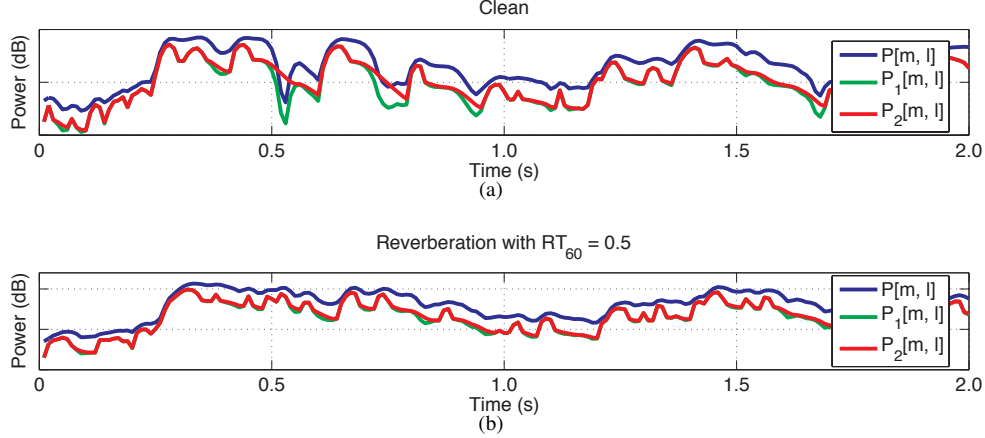


Figure 2: Power contour $P[m, l]$, $P_1[m, l]$ (processed by SSF Type-I processing), and $P_2[m, l]$ (processed by SSF Type-II processing) for the 10-th channel in clean environment (a) and in the reverberant environment (b).

sults in better sound quality for non-reverberated speech, but the latter results in better speech recognition accuracy in reverberant environments. SSF processing is normally applied to both training and testing data in speech recognition applications.

In speech signal analysis, we typically use a window with duration between 20 and 30 ms. With the SSF algorithm, we observe that windows longer than this length are more appropriate for estimating or compensating for noise components, which is consistent with our observations in previous work (*e.g.* [16, 17, 8]). Nevertheless, even if we use a longer-duration window for noise estimation, we must use a short-duration window for speech feature extraction. After performing frequency-domain processing we use an IFFT and the overlap-add method (OLA) to re-synthesize speech, as in [9]. Feature extraction and subsequent speech recognition can be performed on the re-synthesized speech. We call this general approach the “medium-duration analysis and synthesis approach”.

2. Structure of the SSF algorithm

Figure 1 shows the structure of the SSF algorithm. The input speech signal is pre-emphasized and then multiplied by a medium-duration Hamming window as in [9]. This signal is represented by $x_m[n]$ in Fig. 1 where m denotes the frame index. We use a 50-ms window and 10 ms between frames. After windowing, the FFT is computed and integrated over frequency using gammatone weighting functions to obtain the power $P[m, l]$ in the m^{th} frame and l^{th} frequency band as shown below:

$$P[m, l] = \sum_{k=0}^{N-1} |X[m, e^{j\omega_k}] H_l(e^{j\omega_k})|^2, \quad 0 \leq l \leq L-1 \quad (2)$$

where k is a dummy variable representing the discrete frequency index, and N is the FFT size. The discrete frequencies are $\omega_k = \frac{2\pi k}{N}$. Since we are using a 50-ms window, for 16-kHz audio samples N is 1024. $H_l(e^{j\omega_k})$ is the spectrum of the gammatone filter bank for the l^{th} channel evaluated at frequency index k , and $X[m, e^{j\omega_k}]$ is the short-time spectrum of the speech signal for the m^{th} frame, where $L = 40$ is the total number of gammatone channels. After the SSF processing described below, we perform spectral reshaping and compute the IFFT using OLA to obtain enhanced speech.

3. SSF Type-I and SSF Type-II Processing

In SSF processing we first obtain lowpassed power $M[m, l]$ from each channel:

$$M[m, l] = \lambda M[m-1, l] + (1-\lambda)P[m, l] \quad (3)$$

where λ is a forgetting factor that is adjusted for the bandwidth of the lowpass filter. The processed power is obtained by the following equation:

$$P_1[m, l] = \max(P[m, l] - M[m, l], c_0 P[m, l]) \quad (4)$$

where c_0 is a small fixed coefficient to prevent $P[m, l]$ from becoming negative. In our experiments we find that $c_0 = 0.01$ is appropriate for suppression purposes. As is obvious from (4), $P_1[m, l]$ is essentially a highpass filter signal, since the lowpassed power $M[m, l]$ is subtracted from the original signal power $P[m, l]$. From (4), we observe that if power $P[m, l]$ is larger than $M[m, l] + c_0 P[m, l]$ then, $P_1[m, l]$ is the highpass filter output. However, if $P[m, l]$ is smaller than the latter, the power is suppressed. These operations have the effect of suppressing the falling edge of the power contour. We call processing using (4) SSF Type-I.

A similar approach uses the following equation instead of (4):

$$P_2[m, l] = \max(P[m, l] - M[m, l], c_0 M[m, l]) \quad (5)$$

We call this processing SSF Type-II.

The only difference between (4) and (5) is one term, but as shown in Fig 3 and 4, this term has a major impact on recognition accuracy in reverberant environments. We also note that using SSF Type-I processing, if $0.2 \leq \lambda \leq 0.4$, substantial improvements are observed for clean speech compared to baseline processing. In the power contour of Fig. 2, we observe that if we use SSF Type-II, the falling edge is smoothed (since $M[m, l]$ is basically a lowpass signal), which significantly reduces spectral distortion between clean and reverberant environments.

Fig. 3 shows the dependence of performance on the forgetting factor λ and the window length. For additive noise, a window length of 75 or 100 ms provided the best performance. However, for reverberation, 50 ms provided the best performance. Thus we use $\lambda = 0.4$ and a window length of 50 ms.

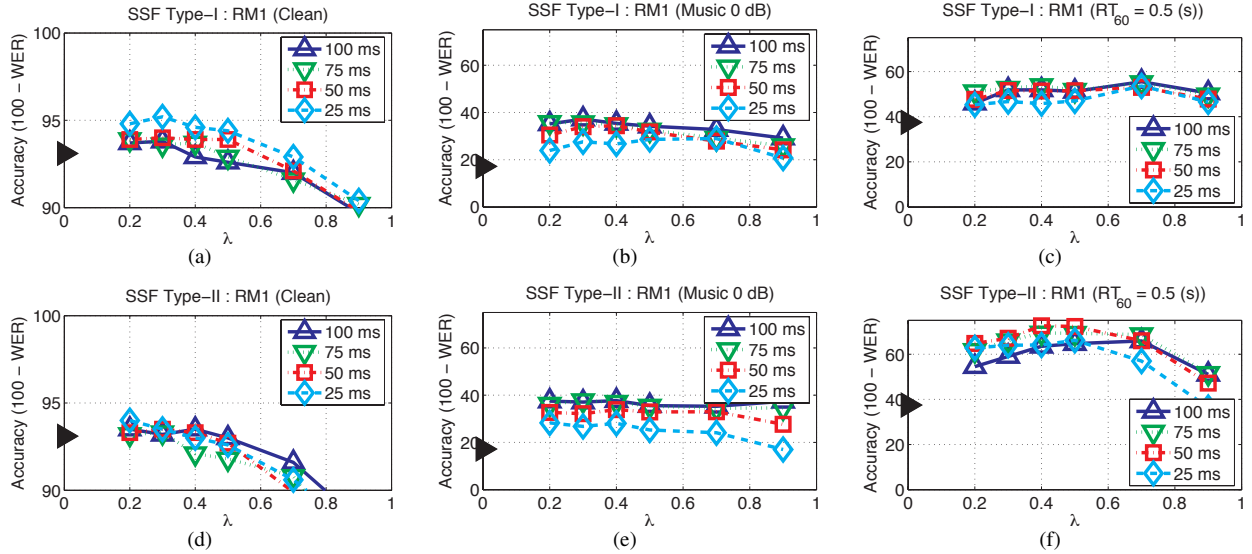


Figure 3: The dependence of speech recognition accuracy on the forgetting factor λ and the window length. In (a), (b), and (c), we used Eq. (4) for normalization. In (d), (e), and (f), we used Eq. (5) for normalization. The filled triangles along the vertical axis represent the baseline MFCC performance in the same environment.

4. Spectral reshaping

After obtaining processed power $\tilde{P}[m, l]$ [which is either $P_1[m, l]$ in (4) or $P_2[m, l]$ (5)], we obtain a processed spectrum $\tilde{X}[m, e^{j\omega_k}]$. To achieve this goal, we use a similar spectral reshaping approach as in [9] and [17]. Assuming that the phases of the original and the processed spectra are identical, we modify only the magnitude spectrum.

First, for each time-frequency bin, we obtain the weighting coefficient $w[m, l]$ as a ratio of the processed power $\tilde{P}[m, l]$ to $P[m, l]$.

$$w[m, l] = \frac{\tilde{P}[m, l]}{P[m, l]}, \quad 0 \leq l \leq L - 1 \quad (6)$$

Each of these channels is associated with H_l , the frequency response of one of a set of gammatone filters with center frequencies distributed according to the Equivalent Rectangular Bandwidth (ERB) scale [18]. The final spectral weighting $\mu[m, k]$ is obtained using the above weight $w[m, l]$

$$\mu[m, k] = \frac{\sum_{l=0}^{L-1} w[m, l] |H_l(e^{j\omega_k})|}{\sum_{l=0}^{L-1} |H_l(e^{j\omega_k})|}, \quad 0 \leq k \leq N/2 - 1, 0 \leq l \leq L - 1 \quad (7)$$

After obtaining $\mu[m, k]$ for the lower half frequency region $0 \leq k \leq N/2$, we can obtain the upper half from the symmetric characteristic:

$$\mu[m, k] = \mu[m, N - k], \quad N/2 \leq k \leq N - 1 \quad (8)$$

Using $\mu[m, k]$, the reconstructed spectrum is obtained by:

$$\tilde{X}[m, e^{j\omega_k}] = \mu[m, k] X[m, e^{j\omega_k}], \quad 0 \leq k \leq N - 1 \quad (9)$$

The enhanced speech $\hat{x}[n]$ is re-synthesized using the IFFT and the OLA method as described above.

5. Experimental results

In this section we describe experimental results obtained on the DARPA Resource Management (RM) database using the SSF algorithm. For quantitative evaluation of SSF we used 1,600 utterances from the DARPA Resource Management (RM) database for training and 600 utterances for testing. We used SphinxTrain 1.0 for training the acoustic models, and Sphinx 3.8 for decoding. For feature extraction we used sphinx_fe which is included in sphinxbase 0.4.1. Even though SSF was developed for reverberant environments, we also conducted experiments in additive noise as well. In Fig. 4(a), we used test utterances corrupted by additive white Gaussian noise, and in Fig. 4(b), we used test utterances corrupted by musical segments of the DARPA Hub 4 Broadcast News database.

We prefer to characterize improvement as the amount by which curves depicting WER as a function of SNR shift laterally when processing is applied. We refer to this statistic as the “threshold shift”. As shown in these figures, SSF provides 8-dB threshold shifts for white noise and 3.5-dB shifts for background music. Obtaining improvements in the presence of background music is generally not easy. For comparison, we also obtained similar results using a modern noise compensation algorithm, vector Taylor series (VTS) [2]. We also conducted experiments using an open source implementation of RASTA-PLP [19]. For white noise, VTS and SSF provide almost the same performance, but for background music, SSF provides a significantly lower word error rate (WER). In additive noise, both SSF Type-I and SSF Type-II provide almost the same WER. For clean utterances, SSF Type-I performs slightly better than SSF Type-II.

To simulate the effects of room reverberation, we used the software package Room Impulse Response (RIR) [20]. We assumed a room of dimensions of 5 x 4 x 3 m, a distance between the microphone and the speaker of 2 m, with the microphones located at the center of the room. In reverberant environments, as shown in Fig. 4(c), SSF Type-II provides the lowest WER by

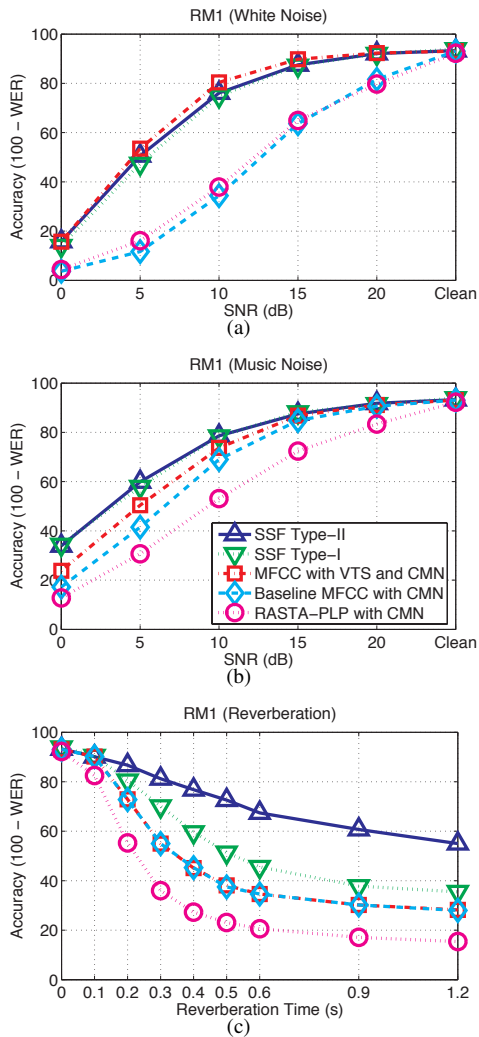


Figure 4: Speech recognition accuracy using various algorithms for speech in (a) white noise (b) musical noise, and (c) under reverberant environments.

a very large margin. SSF Type-I provides the next best WER, but the performance difference between SSF Type-I and SSF Type-II is large. On the contrary, VTS does not provide any improvement in WER, and PLP-RASTA provides worse performance than MFCC.

6. Conclusions

In this paper, we present a new algorithm that is especially robust with respect to reverberation. Motivated by modulation frequency concepts and the precedence effect, we apply first-order high-pass filtering to power coefficients, and the falling edges of power contours are suppressed in two different ways. We observe that the use of a lowpassed signal for the falling edge is especially helpful for reducing spectral distortion in reverberant environments. Experimental results show that this approach is much more effective than the other algorithms considered, and especially in reverberant environments.

Open source MATLAB code for the SSF algorithm may be found at http://www.cs.cmu.edu/~robust/archive/algorithms/SSF_IS2010/. This code was used to obtain the results in Sec. 5.

7. References

- [1] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Nov. 2001, pp. 21–24.
- [2] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, May. 1996, pp. 733–736.
- [3] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 2, Apr. 1997, pp. 851–854.
- [4] C. Kim, Y.-H. Chiu, and R. M. Stern, "Physiologically-motivated synchrony-based processing for robust automatic speech recognition," in *INTERSPEECH-2006*, Sept. 2006, pp. 1975–1978.
- [5] B. Raj and R. M. Stern, "Missing-feature methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, Sept. 2005.
- [6] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing on speech recognition," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.
- [7] —, "Effect of reducing slow temporal modulations on speech recognition," *J. Acoust. Soc. Am.*, vol. 95, no. 5, pp. 2670–2680, May 1994.
- [8] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4574–4577.
- [9] —, "Power function-based power distribution normalization algorithm for robust speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 188–193.
- [10] C. Kim, K. Kumar and R. M. Stern, "Robust speech recognition using small power boosting algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 243–248.
- [11] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, Aug. 1998.
- [12] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," in *EUROSPEECH '91*, Sept. 1991, pp. 413–416.
- [13] P. M. Zurek, *The precedence effect*. New York, NY: Springer-Verlag, 1987, ch. 4, pp. 85–105.
- [14] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1997.
- [15] Y. Park and H. Park, "Non-stationary sound source localization based on zero crossings with the detection of onset intervals," *IE-ICE Electronics Express*, vol. 5, no. 24, pp. 1054–1060, 2008.
- [16] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH-2009*, Sept. 2009, pp. 28–31.
- [17] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *INTERSPEECH-2009*, Sept. 2009, pp. 2495–2498.
- [18] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica - Acta Acustica*, vol. 82, no. 2, pp. 335–345, 1996.
- [19] D. Ellis. (2006) PLP and RASTA (and MFCC, and inversion) in matlab using melfcc.m and invmelfcc.m. [Online]. Available: <http://labrosa.ee.columbia.edu/matlab/rastamat/>
- [20] S. G. McGovern, "A model for room acoustics," <http://2pi.us/rir.html>.