interspeech 2008 incorporating SST 08

Effect of Compressing the Dynamic Range of the Power Spectrum in Modulation Filtering Based Speech Enhancement

James G. Lyons and Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane Queensland 4111, Australia
j.lyons@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In the modulation-filtering based speech enhancement method, noise suppression is achieved by bandpass filtering the temporal trajectories of the power spectrum. In the literature, some authors use the power spectrum directly for modulation filtering, while others use different compression functions for reducing the dynamic range of the power spectrum prior to its modulation filtering. This paper compares systematically different dynamic range compression functions applied to the power spectrum for speech enhancement. Subjective listening tests and objective measures are used to evaluate the quality as well as the intelligibility of the enhanced speech. The quality is measured objectively in terms of the Perceptual Estimation of Speech Quality (PESQ) measure and the intelligibility in terms of the Speech Transmission Index (STI) measure. It is found that $P^{0.3333}$ (power spectrum raised to power $1/3$) results in the highest speech quality and intelligibility.

Index Terms: modulation spectrum, modulation filtering, speech enhancement

1. Introduction

Many speech enhancement methods reported in the literature [1] use the short-time Fourier analysis modification synthesis framework, where the magnitude (or power) spectrum is modified to suppress the noise distortion, while the phase spectrum is left unchanged. The modulation filtering based speech enhancement method, recently proposed in [2], can also be implemented in this framework, where the modified power spectrum is obtained by filtering the temporal trajectories of the power spectrum. This process is known as modulation filtering and is used in other speech processing applications (such as speech and speaker recognition) as well.

The modulation spectrum is defined as the Fourier transform of the time trajectories of individual frequency components of the power spectrum (or its nonlinearly compressed version). The importance of different modulation frequencies for speech intelligibility has been investigated by many researchers in the literature. For example, Drullman et al. [3] have reported a study where the speech signal is split into a number of frequency subbands, the temporal envelope of each subband is lowpass filtered, and the original carriers and filtered envelopes of all the subbands are combined to reconstruct the output speech signal. Using these reconstructed signals, they have shown that modulation frequencies below 16 Hz are important for intelligibility. In a similar study carried out using highpass modulation filters, Drullman et al. [4] have shown the modulation frequencies above 4 Hz are important for intelligibility. Arai et al. [5] have applied filters to the time trajectories of LPC cepstrum, showing that applying a bandpass filter with passband between 1 Hz and 16 Hz does not impair speech intelligibility. They have also shown that some modulation frequencies are more important than others, with the region around 4 Hz being the most important for intelligibility.

Since the region of the modulation spectrum between 1 and 16 Hz contributes the most to intelligibility, the information outside this region can be removed for the purposes of speech enhancement. Hermansky et al. [6, 7] have used this concept to propose a speech enhancement procedure where FIR filters are applied to the time trajectories of the cubic root compressed short-time power spectrum to achieve better speech quality in the presence of additive noise. Falk et al. [8] have applied bandpass filtering to the temporal trajectories of short-time magnitude spectrum to achieve enhancement. Fujioka et al. [9] have used 1 to 16 Hz modulation filtering of the time trajectories of the short-time power spectrum (i.e., without any compression) for speech enhancement.

When low frequency information is removed from the time trajectories of the power spectrum, the resulting filtered spectrum may drop below zero. In this instance, half-wave rectification is usually employed to zero any negative portions. This half-wave rectification, however, results in distortion of the speech signal. Falk et al. overcomes this problem by estimating the "speech only" low frequency modulation content from the speech dominated bandpass modulation content of the noisy signal, then combining the two [8].

Modulation filtering has also been used to improve robustness of automatic speech recognition systems, with Nadeu et al. showing that filtering time sequences of spectral parameters can improve the recognition rate, and that the most important part of the modulation spectrum lies at around 3 Hz [10]. Hirsch, Meyer and Ruehl have demonstrated that highpass filtering of the subband envelopes can improve speech recognition accuracy in additive noise [11]. Hermansky and Morgan [2] have shown that filtering of the log-power spectrum before feature extraction results in better robustness to channel distortion. Hermansky and Morgan have also filtered a Lin-Log compressed power spectrum and shown that this results in increased robustness in additive noise.

It can be noted that many of the papers reviewed here use different compression functions for reducing the dynamic range of the power spectrum before modulation filtering is applied. The aim of this paper is to determine which power spectrum compression function results in the best enhanced speech, measured using objective speech quality and intelligibility measures as well as through subjective listening tests.

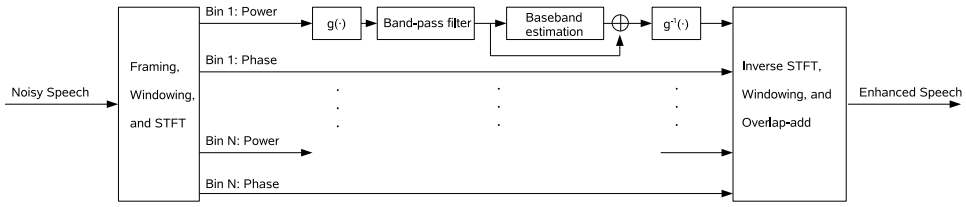

Figure 1: Block diagram of Modulation filtering based speech enhancement method

2. Modulation filtering based speech enhancement method

The modulation filtering based speech enhancement method is implemented here using the short-time analysis-modification-synthesis framework as shown in Fig. 1. For this, the input speech signal $x(t)$ is analysed frame-wise with 20 ms Hamming window and 10 ms frame-shift. Let $X_k(f)$ be the short-time Fourier transform (computed through 512 point FFT) for the f^{th} frequency bin in the k^{th} frame. We define the short-time power spectrum of speech as: $P_k(f) = |X_k(f)|^2$. Let N be the number of FFT bins and K be the number of frames in the noisy speech utterance to be enhanced, then $P_k(f), k = 1 \ldots K$, denotes the time trajectory of the f^{th} frequency bin for the given utterance.

In order to show the importance of modulation filtering, we compute the modulation spectrum of a 128 ms section of vowel extracted from a speech utterance. We consider a frequency bin f centered around 800 Hz and use the log power spectrum of the time trajectory of $[P_k(f)]^{0.5}$ to represent the magnitude of the modulation spectrum for this bin. Figure 2 shows this modulation spectrum for the vowel section. We also show in this figure the corresponding modulation spectra for the signal corrupted by additive white noise distortion at 15 dB and 5 dB signal-to-noise ratios (SNRs). This figure shows that the modulation spectra are almost identical for all cases below 20 Hz, which is in accordance with previous literature. The region of the modulation spectrum after 20 Hz is heavily influenced by the addition of noise, with the 5 dB case having substantially more energy in that region. This justifies the use of lowpass modulation filters with cutoffs at roughly this frequency.

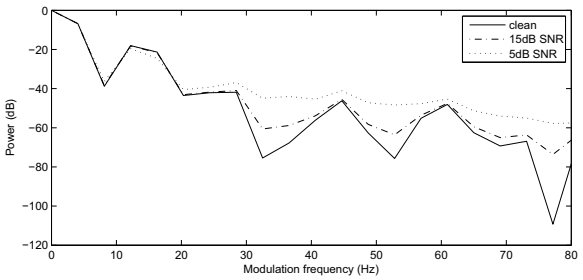

Figure 2: Modulation spectrum (Log power spectrum of time trajectory of short-time magnitude spectrum) at an FFT bin around 800 Hz.

Modulation filtering is performed by applying a bandpass filter to time trajectory of $g(P_k(f))$, $k = 1 \ldots K$ for $f = 1 \ldots N$. Here, we use a 301-length FIR filter designed using the Parks-Mclellan method with a passband between 1 and 16 Hz and $g(\cdot)$ is a compression function that is applied to the power spectrum to reduce its dynamic range. For example, $g(P_k(f)) = (P_k(f))^{0.3333}$ represents the cubic root compression of the power spectrum. Bandpass filtering produces bandpass components of the compressed temporal envelope. The baseband component of the compressed temporal envelope is estimated from the bandpass component. The estimated baseband component and the bandpass component are added to form an estimate of the compressed temporal envelope. This estimated envelope may have some negative values. To remove these negative values, half-wave rectification is performed. The inverse compression function, $g^{-1}(\cdot)$, is then applied to undo the effect of compression. The resulting estimated envelope is used to produce the short-time magnitude spectrum. This modified magnitude spectrum is used with the original short-time phase spectrum to reconstruct the enhanced signal through inverse FFT, windowing and overlap-add synthesis.

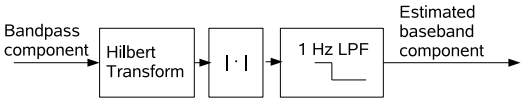

Figure 3: Block diagram of 'clean' base-band estimation from bandpass filtered time trajectory of noisy spectrum. This illustrates the Baseband Estimation step in Fig. 1

The baseband estimation procedure shown in Fig. 3 is from [8]. It computes the absolute value of the Hilbert transform of the bandpass component of the compressed temporal envelope. The resulting envelope is lowpass filtered with 1 Hz cutoff frequency to form an estimate of baseband component.

3. Experimental Results

3.1. Objective Evaluation Methods

In this subsection, we evaluate the quality and intelligibility of speech using objective measures. In our evaluation, we compute objective scores over a subset of the TIMIT database.

For speech quality evaluation purposes, we employ an objective speech quality measure, namely the Perceptual Estimation of Speech Quality (PESQ)[1]. PESQ prediction maps Mean Opinion Scores (MOS) to a range between -0.5 and 4.5, where 1.0 corresponds to bad and 4.5 corresponds to distortionless.

For speech intelligibility evaluation, we employ the Speech Transmission Index (STI) measure. The STI algorithm [12, 13] measures the extent to which speech envelope modulations are preserved in degraded listening environments. STI scores range from 0 (completely unintelligible) to 1 (perfect intelligibility).

The STI algorithm we employ here uses the speech signal as the probe.

Different compression functions (function $g(\cdot)$ in figure 1) are used during speech enhancement. The objective speech quality results are shown in table 1, and the objective speech intelligibility results are shown in table 2.

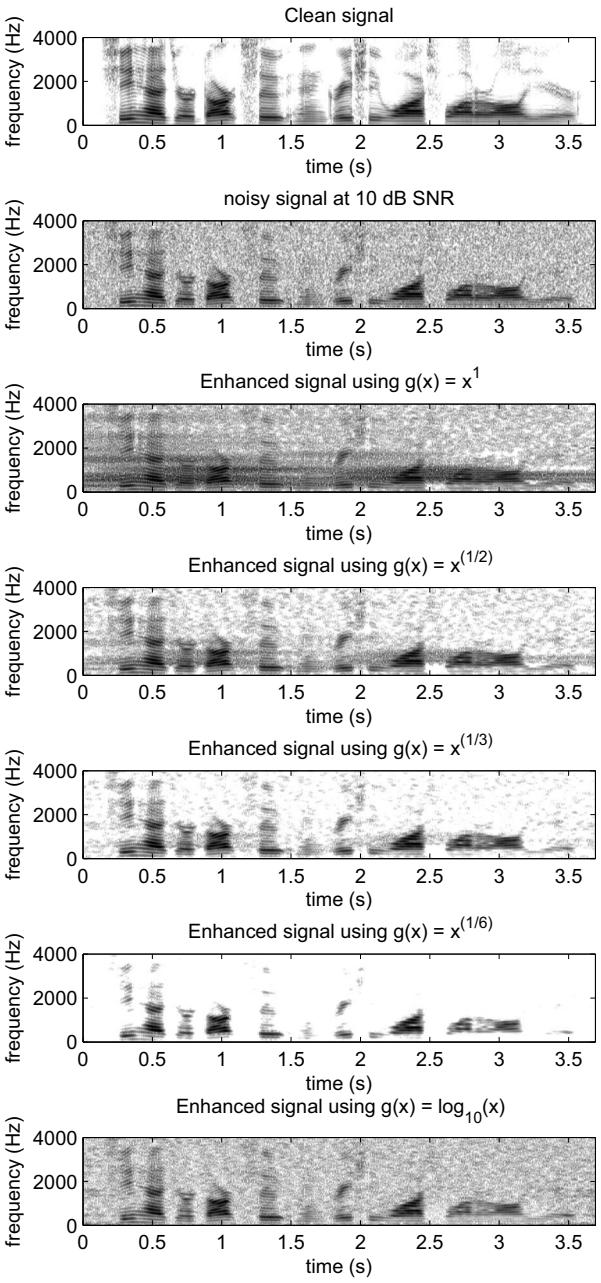

Figure 4: Spectrograms of original, noisy and enhanced speech. Noisy signal is corrupted with 10 dB SNR white gaussian noise. Enhanced spectrograms show the effect of different compression functions.

From table 1, it is evident that the function $P^{0.1667}$ gives the best speech quality performance. The function P^1 (uncompressed power spectrum) performs quite poorly. This is because the lowpass filtered Hilbert transform is not a good estimator of the baseband component for this case. The stronger compression functions mitigate this effect. It is also evident that the log compression performs poorly for additive noise.

Table 2 shows the STI scores for each power spectrum compression function at several SNRs. The trend that can be seen from these scores is identical to the PESQ scores, with the $P^{0.1667}$ compression function performing the best.

3.2. Subjective Listening tests for speech quality

For listening tests, we used 5 listeners and two utterances (one from a male speaker and other from a female speaker). Each utterance was corrupted with white gaussian noise at 10 dB SNR before being enhanced using each of the compression functions tested earlier. For each enhancement type, listeners listened to the original signal followed by one of the corrupted signals. Each corrupted signal was scored by giving it a number between 1 and 5, with 5 corresponding to imperceptible distortion and 1 corresponding to very annoying. The listeners were allowed to listen to the original and each of the enhanced utterances as many times as they needed.

The results, listed in table 3, show that filtering of the $P^{0.1667}$ compressed power spectrum results in the least annoying residual noise, agreeing with the objective speech quality results. Filtering the uncompressed power spectrum results in serious degradation, and it was felt by listeners that the resulting signal was even worse than the original noisy signal in some cases. Filtering the log compressed power spectrum was not considered to significantly improve the speech quality in the presence of additive noise.

Informal discussions with the listeners revealed that the stronger compression functions ($P^{0.3333}$ and $P^{0.1667}$) had less annoying noise than the less aggressive compression functions, but that the speech started to sound more 'bottled' for the $P^{0.1667}$ case, indicating some loss of speech information.

3.3. Subjective Listening tests for intelligibility

Listening tests were performed to determine the intelligibility of the enhanced speech. This was tested with -5dB SNR white gaussian noise with a single listener. Six commonly occurring stop consonants in Australian English in aCa context were recorded, spoken in a carrier sentence 'Hear aCa now'. For example, for the consonant /d/, the recorded utterance is 'Hear ada now'. These 6 consonants in the carrier sentence are recorded for four speakers: two males and two females. The recordings are sampled at 16 kHz (16-bit). Each recording is corrupted, then an enhanced version is created using each of the power spectrum compression functions applied before modulation filtering. The clean, noisy and enhanced signals were played in random order, with the task being to identify each utterance as one of the 6 consonants b,d,g,k,p,t. The results are summarized in table 4.

The results of the intelligibility tests show that the consonant identification accuracy is best with the $P^{0.333}$ power spectrum compression function. In the speech quality tests performed earlier, we have shown that the $P^{0.1667}$ compression function results in the least 'annoying' residual noise. However, the intelligibility tests show that too much speech information is lost in this case, though the noise might be suppressed better. The cube-root compression results in the best quality enhanced speech without losing significant amounts of speech information. It is also interesting to note that modulation filtering of the cube-root compressed power spectrum results in improved intelligibility compared to the noisy case.

Table 1: Results of experiments showing PESQ of enhanced speech with different compression functions.

	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	25 dB	30 dB
P^1	1.416	1.643	1.851	1.975	2.033	2.072	2.097	2.114
$P^{0.5}$	1.567	1.881	2.181	2.404	2.550	2.636	2.690	2.727
$P^{0.3333}$	1.644	2.002	2.346	2.652	2.876	3.029	3.138	3.214
$P^{0.1667}$	1.654	2.035	2.360	2.669	2.933	3.123	3.256	3.328
$log_{10}(P)$	1.387	1.615	1.873	2.151	2.431	2.678	2.878	3.002
unenhanced	1.355	1.593	1.888	2.221	2.575	2.924	3.266	3.599

Table 2: Results of experiments showing STI of enhanced speech with different compression functions.

	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	25 dB	30 dB
P^1	0.391	0.455	0.514	0.554	0.572	0.578	0.581	0.582
$P^{0.5}$	0.562	0.676	0.778	0.846	0.876	0.882	0.879	0.876
$P^{0.3333}$	0.681	0.801	0.907	0.959	0.974	0.974	0.971	0.968
$P^{0.1667}$	0.846	0.925	0.971	0.986	0.989	0.989	0.988	0.987
$log_{10}(P)$	0.438	0.552	0.684	0.801	0.887	0.931	0.946	0.949
unenhanced	0.361	0.474	0.614	0.751	0.863	0.944	0.987	0.998

Table 3: Results of listening tests MOS scores of enhanced speech with different compression functions.

	clean	noisy	P^1	$P^{0.5}$	$P^{0.333}$	$P^{0.167}$	$log_{10}(P)$
10 dB SNR	5	2.85	1.52	2.19	2.86	3.22	2.61

Table 4: Results of listening tests showing consonant recognition accuracy. Noisy speech is corrupted with -5 dB SNR white gaussian noise. All measurements are in per cent (%) of consonants correctly identified.

	clean	noisy	P^1	$P^{0.5}$	$P^{0.333}$	$P^{0.167}$	$log_{10}(P)$
-5 dB SNR	100	87.5	87.5	91.67	95.83	70.83	75

4. Conclusion

In this paper, we have studied the modulation-based speech enhancement method and investigated the role of the compression function applied to the short-time power spectrum before modulation filtering. In our experiments, we have used objective as well as subjective measurements to measure the quality and intelligibility of the enhanced speech. For objective measurements, PESQ is used for speech quality and STI for speech intelligibility. For subjective measurements, human listening tests are used.

It has been shown that the cube-root compressed power spectrum provides the best quality enhanced speech while providing good intelligibility. Stronger compression functions such as sixth-root of power spectrum result in over suppression of noise resulting in a reduction of speech intelligibility. Weaker compression functions such as using the power spectrum or square root of the power spectrum result in insufficient noise removal. Filtering the log compressed power spectrum does not result in significant improvements in the presence of additive noise.

5. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[2] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[3] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *JASA*, vol. 95, pp. 2670–2680, 1994.

[4] ——, "Effect of reducing slow temporal modulations on speech reception," *JASA*, vol. 95, pp. 2670–2680, 1994.

[5] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. ICSLP*, 1996, pp. 2490–2493.

[6] H. Hermansky, E. A. Wan, and C. Avendano, "Speech enhancement based on temporal processing," in *Proc. ICASSP*, 1995, pp. 405–408.

[7] H. Hermansky, E. Wan, and C. Avendao, "Noise suppression in cellular communications," in *2nd IEEE Workshop IVTTA*, 1994, pp. 85–85.

[8] T. Falk, S. Stadler, W. B. Kleijn, and G. Chan, "Noise suppression based on extending a speech-dominated modulation band," in *Proc. ICSLP*, 2007, pp. 970–973.

[9] K. Fujioka, N. Hayasaka, Y. Miyanaga, and N. Yoshida, "Noise reduction of speech signals by running spectrum filtering," *Syst. Comput. Japan*, vol. 37, no. 14, pp. 52–61, 2006.

[10] C. Nadeu, P. Pach'es-Leal, and B. Juang, "Filtering of time sequences of spectral parameters for speech recognition," in *Speech Communication*, vol. 22, 1997, pp. 315–332.

[11] H. G. Hirsch, P. Meyer, and H. W. Ruhl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. Eurospeech*, 1991, p. 413.

[12] T. Houtgast and H. J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *JASA*, vol. 77, pp. 1069–1077, 1985.

[13] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *JASA*, vol. 116, pp. 3679–3689, 2004.