TARGET TALKER ENHANCEMENT IN HEARING DEVICES

Steven M. Schimmel[1,2]

Laboratory for Experimental Audiology
University of Zurich
CH-8091 Zurich, Switzerland

Les E. Atlas[2]

Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA

ABSTRACT

We describe a novel coherent modulation filtering technique for single channel target talker enhancement in the presence of interfering talkers. For this technique, we have expanded our previous work on coherent modulation filtering with a carrier estimator that is more robust to speech from interfering talkers, and a modulation filter that operates on shorter time-scales. We have evaluated the technique in a subjective listening test, which indicates that the novel target talker enhancement technique achieves a moderate improvement in speech reception. We summarize our observations on single channel target talker enhancement and conclude with directions for further research.

***Index Terms*—** Modulation filtering, speech enhancement, carrier estimation, acoustic resonator filters, hearing aids

1. INTRODUCTION

Most cochlear implant users and users of hearing aids have great difficulty to focus on a target talker in the presence of interfering talkers, a condition that typically occurs in bars, restaurants and other social settings. Modern hearing instruments address this "cocktail party" problem in several ways [1]. They use directional microphones or microphone arrays to enhance speech from the front, thus reducing interference from speakers from other directions, and adaptive (single channel) noise suppression techniques to reduce background noise. Despite the fact that these methods improve the SNR and can reduce the listening stress, they have yet to prove that they enhance speech intelligibility [1].

In this paper, we study the use of coherent modulation filtering to enhance a target talker in the presence of one or more interfering talkers. This choice is motivated by psychoacoustic studies such as [2] that support the belief that the auditory system analyzes and perhaps even segregates sounds in the modulation frequency domain. Moreover, our own earlier work on modulation filtering for talker separation and musical instrument separation showed promising results for this approach [3, 4]. Finally, a modulation filtering approach to target talker enhancement is complementary to other approaches such as directional microphones.

Coherent modulation filtering is a class of non-linear signal processing techniques that can be described by the framework depicted in Fig. 1. A broadband signal $x(n)$ is separated into complex-valued subbands $x_k(n)$. A carrier estimator, denoted $\mathfrak{D}^c$, decomposes each subband into a complex-valued low-frequency modulator $m_k(n)$ and a complex-valued narrowband carrier $c_k(n)$, such that $x_k(n) = m_k(n)c_k(n)$ and $|c_k(n)| = 1$. The modulator is filtered by the filter $g_k(n)$ and recombined with the (unmodified) carrier. A broadband modulation-filtered signal $\tilde{x}(n)$ is reconstructed by summing the modified subbands $\tilde{x}_k(n)$. Coherent modulation filtering techniques, and their merit over other modulation filtering techniques, are discussed in detail in [3, 5, 6].

To successfully apply coherent modulation filtering to the problem of target talker enhancement required us to make several modifications to the conventional implementation of coherent modulation filters. For example, previously defined carrier estimators for coherent modulation filtering (e.g., [3, 5]) are designed to estimate the subband carriers of speech from a single talker in quiet. In the presence of an interfering talker, however, their subband carrier estimates are biased by the interfering talker's carriers. Therefore, the carrier estimator had to be redesigned to be more robust to interfering speech. Furthermore, most existing coherent modulation filtering systems operate on time-scales in the order of 50–250 ms. However, in order to manipulate a target talker independently of interfering talkers, we found that the modulation filter needed to operate on much smaller time-scales of approximately 5–25 ms. Smaller time-scales are also essential for a low-latency implementation of the processing algorithm such that it is suitable for hearing devices.

Details of the redesigned components of the coherent modulation filtering technique for target talker enhancement are given in section 2. The subjective listening test used to evaluate the technique is described in section 3, and the test results are given in section 4. Conclusions and a discussion are presented in section 5.

2. METHOD

The novel carrier estimator is based on a target fundamental frequency estimator that detects and estimates the target talker's fundamental frequency as a function of time even in the presence of one or more interfering talkers. The novel modulation filter is implemented as a bank of resonator filters to increase its precision in time. A mixing stage has also been added as a postprocessing step to the modulation filter, in order to minimize artifacts and distortion in the

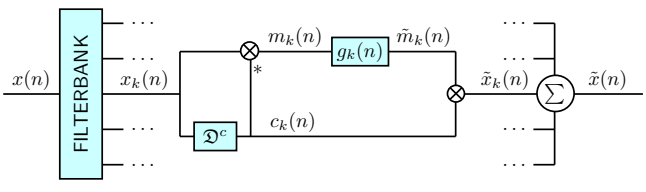

Fig. 1: Coherent modulation filtering framework.

[1]The first author performed the work while at the Department of Electrical Engineering of the University of Washington.

[2]This research was partially supported by AFOSR Grant FA95500610191.

output. Each of these components is described in more detail in the following sections.

2.1. Target fundamental frequency estimator

The target fundamental frequency estimator detects harmonic structures in the input signal over the fundamental frequency range of the target talker, $\omega = [\omega_l, \omega_h]$, which is assumed to be known *a priori*. The estimator is based on a modified version of the *harmonic product spectrum* (HPS) method by Schroeder [7]. It computes the discrete short-time Fourier transform $X(n,\omega)$ of the input signal $x(n)$,

$$X(n,\omega) = \sum_m x(m)w(n-m)e^{-j\omega m}, \quad (1)$$

and sums the log-magnitude-squared of $X(n,\omega)$ over the first P harmonic multiples of the target talker's fundamental frequency range,

$$D(n,\omega) = \sum_{p=1}^{P} \log|X(n,p\omega)|^2, \quad \omega_l \le \omega \le \omega_h. \quad (2)$$

The HPS frame $D_n(\omega) = D(n,\omega)$ represents the total amount of energy that is present in the first P harmonics of the frequency ω at time n. Signal components that are harmonically related add up constructively and create a maximum in $D_n(\omega)$.

The HPS exhibits several desirable features for carrier estimation. First, it avoids octave errors in the fundamental frequency estimate, because typically $\omega_h < 2\omega_l$. Furthermore, the HPS is robust to noise because it uses all of a speech signal's strongest harmonics to estimate its fundamental frequency. Finally, it allows a simple and efficient implementation, for example via quadratic interpolation of frequency samples [8, 9].

Based on the HPS, a target voicing activity detector $v(n)$ is recursively defined by

$$v(n) = \begin{cases} 1, & \text{if } v(n-1)=0,\ p(n)>p_1,\ q(n)>q_1 \\ 0, & \text{if } v(n-1)=1,\ p(n)<p_0,\ q(n)<q_0 \\ v(n-1), & \text{otherwise,} \end{cases} \quad (3)$$

where $p(n)$ and $q(n)$ are two empirical measures that express the "peakedness" of the maximum of $D_n(\omega)$ with respect to the (other) local maxima of $D_n(\omega)$ and to the entire frame $D_n(\omega)$, respectively. They are computed according to

$$p(n) = \frac{\max[D_n(\omega)] - \mu_l(n)}{\sigma_l(n)}, \quad (4)$$

$$q(n) = \frac{\max[D_n(\omega)] - \mu(n)}{\sigma(n)}, \quad (5)$$

where $\mu_l(n)$ and $\sigma_l^2(n)$ are the mean and variance of the local maxima of the frame $D_n(\omega)$, not including the global maximum itself, and $\mu(n)$ and $\sigma^2(n)$ are the mean and variance of the entire frame $D_n(\omega)$. The parameters p_0, p_1, q_0, and q_1 in equation (3) are empirically determined minimum and maximum thresholds on the peakedness of HPS frames.

Given the target voicing activity detector, the target fundamental frequency $f(n)$ is defined by

$$f(n) = \begin{cases} \text{argmax}_\omega\, D_n(\omega), & v(n)=1 \\ 0, & v(n)=0 \end{cases}, \quad (6)$$

where $f(n)=0$ indicates that the target voice was not detected and no fundamental frequency estimate could be made. The subband carrier estimates necessary for modulation filtering are defined in turn by $f_k(n) = kf(n)$ for $k = 1, \ldots, K$.

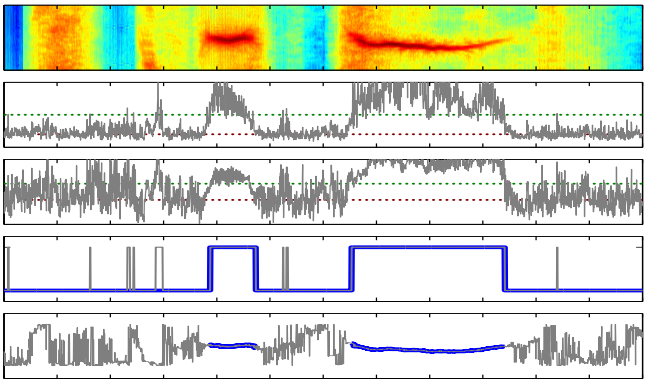

Fig. 2: Components of the target fundamental frequency estimator. From top to bottom: Harmonic product spectrum $D(n,\omega)$; peakedness measures $p(n)$ and $q(n)$, with minimum and maximum threshold shown in red and green; target voicing activity detection $v(n)$, shown before (gray) and after (blue) suppressing short bursts; fundamental frequency estimate $f(n)$, shown as $\text{argmax}_\omega\, D_n(\omega)$ (gray) and final estimate (blue).

The target fundamental frequency estimator employs a few subtle heuristics, e.g. suppressing short bursts in voicing activity for the continuity and smoothness of the fundamental frequency estimate. Details of these heuristics are omitted here due to space constraints, but are given in [6]. The operation of the fundamental frequency estimator is illustrated by an example of its components in Fig. 2.

2.2. Modulation filter

The modulation filter of our approach to target talker enhancement is implemented as a bank of time-varying second order IIR resonator filters. Each resonator filter can be interpreted as a lowpass modulation filter in a time-varying subband, where each subband is centered on a harmonic of the speech signal.

The idea of time-varying resonator filters is similar to the dynamic tracking filter originally proposed in [10] for satellite communication systems, which was later redefined and extended in [11] for speech signals. However, our implementation of the bank of resonator filters differs from such dynamic tracking filters in an important way: we separate the tracking of a harmonic from filtering it. This allows us to use the target fundamental frequency estimator described in section 2.1, which achieves greater noise robustness by exploiting the harmonic structure of voiced speech to track the fundamental frequency, rather than tracking each harmonic independently.

The k-th resonator in the bank of filters is defined by the time-varying difference equation

$$y_k(n) = 2\gamma_k \cos(f_k(n))y_k(n-1) - (2\gamma_k - 1)y_k(n-2) + (1-\gamma_k)[x(n) - x(n-2)]. \quad (7)$$

The gain term, $\gamma_k = \frac{1}{1+\beta_k}$, ensures that the filter has unit response at the resonance frequency. It is defined in terms of the bandwidth factor, $\beta_k = \frac{1}{3}\sqrt{3}\tan(B_k/2)$, which depends on the resonator band-

width B_k. Given the output of each resonator filter, as defined by (7), the output of the multiresonator filterbank is defined as the sum of the individual resonators, $y(n) = \sum_{k=1}^{K} y_k(n)$. The number of resonators in the filterbank is determined by the lower limit, ω_l, of the target talker's fundamental frequency range, and by the cutoff frequency, ω_{lp}, of the lowpass filter used in the mixing stage. The number K should be large enough such that $K\omega_l > \omega_{lp}$.

Note that the time-varying multiresonator filterbank described here differs from adaptive comb filters (e.g., [12–14]) in two important aspects. First, the resonance frequencies of the resonator filters can take on any value, and can therefore smoothly track speech harmonics, whereas the comb filter frequencies can only be integer subharmonics of the sampling frequency. Second, unlike the constant bandwidth of the comb filter's "combs", the bandwidths of the resonator filters are independent of each other and proportional to their resonance frequency, to accommodate the (typically) greater bandwidth of higher harmonics of speech.

2.3. Mixing

In the mixing stage, the target voicing activity detection, $v(n)$, is combined with the output of the modulation filter, $y(n)$, and mixed with the original signal, $x(n)$, as follows. Both the original signal and the output are filtered with a lowpass filter, $h_{lp}(n)$, resulting in the low-frequency signals

$$x_{lp}(n) = x(n) * h_{lp}(n) \tag{8}$$

$$y_{lp}(n) = y(n) * h_{lp}(n), \tag{9}$$

and their high-frequency counterparts

$$x_{hp}(n) = x(n) - x_{lp}(n) \tag{10}$$

$$y_{hp}(n) = y(n) - y_{lp}(n), \tag{11}$$

where we have assumed for convenience that $h_{lp}(n)$ has zero phase. The lowpass filter's cutoff frequency ω_{lp} is chosen such that the filter suppresses the higher harmonics of the target's speech signal. A typically value for ω_{lp} is in the range of 1500–2000 Hz.

The voicing activity signal $v(n)$ is then used to modulate the lowpass and highpass filtered input and filterbank signals, as follows:

$$\tilde{x}(n) = v(n)\left[\beta_1 y_{lp}(n) + \frac{x_{hp}(n)}{\beta_2}\right] + [1 - v(n)]\frac{x(n)}{\beta_3}. \tag{12}$$

The mixing constant $\beta_1 > 1$ amplifies the low-pass filtered output of the multiresonator filterbank when voiced speech from the target talker is detected. This enhances the harmonic structure of voiced speech from the target talker in low frequencies, while avoiding the "metallic" artifacts commonly associated with an overly forced harmonic structure in high frequencies. The second term of equation (12) controls the amount of high frequencies that are passed from the input $x(n)$ to the output $\tilde{x}(n)$ at times when voiced speech from the target talker is detected. At those times, the signal component $x_{hp}(n)$ contains the higher harmonics of the target talker's speech, and potentially contains high frequencies from the interfering talkers. The mixing constant $\beta_2 > 1$ that attenuates this component is a compromise between attenuating the interfering talker to an acceptable level, while maintaining enough of the target talker's higher harmonics. By passing the higher harmonics of the target talker unfiltered, we have found that much of the naturalness of the target's voiced speech is preserved. The mixing constant $\beta_3 > 1$ attenuates the input signal when no voiced speech from the target talker is detected. It is important for the intelligibility of the target talker to pass this component at a moderate level, because it contains the unvoiced parts of the target's speech signal. It also helps to maintain the overall quality of the signal to pass this component to the output in attenuated form. In informal listening, we found that mixing the input and output signals as described above contributed significantly to speech intelligibility, and produced artifact free and natural sounding signals.

Table 1: Parameter values of the novel coherent modulation filtering algorithm as used in the subjective listening test.

description	parameter	value
target f_0 range	$\omega = [\omega_l, \omega_h]$	$[212, 250]$ Hz
number of harmonics	P	14
peakedness thresholds	$\{p_0, p_1, q_0, q_1\}$	$\{2, 5, 2.5, 5\}$
number of resonators	K	14
resonator bandwidths	B_k	$\frac{1}{2}(3k + 23)$ Hz
lowpass cutoff	ω_{lp}	2000 Hz
mixing constants	$\{\beta_1, \beta_2, \beta_3\}$	$\{2, 2, 2\}$

3. EXPERIMENT

We evaluated the performance of the novel modulation filtering approach to target talker enhancement using a subjective listening test. The objective of the listening test was to measure the speech reception threshold (SRT) of a target talker in two-talker babble under various processing conditions. We performed the listening test on three bilateral hearing loss patients over a hearing aid, and on six normal hearing subjects over a cochlear implant simulation as described in [15]. This last subject group was included as a substitute for cochlear implant users, who we could not recruit in time for this test.

The test stimuli were similar to those in [16]. Each of twelve spondee (two-syllable) words, spoken by the target female talker, was mixed with a two-talker babble noise signal that consisted of two sentences spoken by a male and a female talker different from the target talker. Spondee and babble were mixed at signal-to-noise ratios (SNR) ranging from -50 dB to +20 dB in steps of 2 dB. The RMS amplitude of the spondee was kept constant in all mixtures, and the RMS amplitude of the babble was scaled to the desired SNR. The stimuli were presented in three processing conditions: (1) original stimuli ("unprocessed"); (2) target talker enhancement using the novel coherent modulation filter ("coherent"); and (3) same as 2, but with target talker detection and estimation done on the spondee alone, i.e., without interfering talkers ("coherent in quiet"). This condition was included to evaluate the performance of the modulation filter independent of the target fundamental frequency estimator. For the hearing impaired subjects, all three conditions were presented over a hearing aid with noise reduction and directionality disabled. To compare our target talker enhancement approach to the noise reduction of the hearing aid, we repeated the first condition with the hearing aid's noise reduction enabled as a fourth condition ("noise reduction"). The parameter values of the novel algorithm that were used in the listening test are listed in Table 1. The listening test was setup as an adaptive, twelve alternative forced-choice SRT test using a simple 1-up, 1-down method [17]. It was repeated until 14 reversals in SNR were completed. The mean of the SNR at the last 10 reversals was taken as the estimate of the 50% correct SRT. Each subject completed 6 SRT measurements in all processing conditions. Full details of the listening test are given in [6].

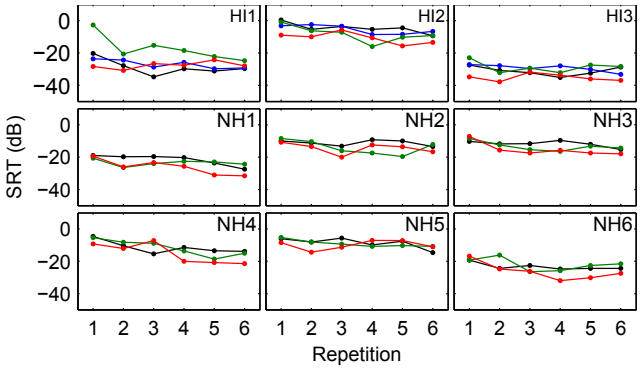

Fig. 3: SRT per processing method for hearing impaired (HI) and normal hearing (NH) subjects as a function of repetition number. (See Table 2 for key.)

4. RESULTS

The results of the listening test are shown in Fig. 3 and summarized in Table 2. Fig. 3 shows the performance of individual subjects for each processing method. A learning effect is visible, as SRTs generally decrease with increasing repetition number. Overall, the hearing impaired subjects received no benefit from hearing aid noise reduction (SRT increased 1.0 dB), and no benefit from coherent processing (SRT increased 3.5 dB). They did however benefit from coherent in quiet processing (SRT decreased 2.9 dB). The effect of the processing method was, however, not significant according to a repeated measures ANOVA. Furthermore, their learning effect was very significant ($p < 0.01$) and the interaction between processing method and repetition number was significant ($p < 0.05$). The normal hearing subjects received benefit from the coherent processing (SRT decreased 1.0 dB) and from the coherent in quiet processing (SRT decreased 3.2 dB). For these subjects, the effect of the processing method as well as the learning effect was very significant ($p < 0.001$), but there was no significant interaction between them.

5. CONCLUSIONS AND DISCUSSION

The subjective listening test indicates that the coherent modulation filtering technique moderately increases speech intelligibility. However, the number of participants in the test is low, and firm conclusions can not be drawn without additional testing. Furthermore, the test shows that the increase in speech intelligibility is greatest for the "coherent in quiet" processing condition, suggesting that the technique's performance is limited by the target talker detector and fundamental frequency estimator. This could be, on one hand, because the detector and estimator are required to function at very low SNRs for the listening test; much lower than they were really designed for, and much lower than what is representative of real "cocktail parties". On the other hand, the estimator could likely be improved by incorporating a more sophisticated algorithm, such as dynamic programming or particle filtering, to track the target's fundamental frequency or harmonics over time. Moreover, the target talker detector uses a straightforward model of the target talker based on its fundamental frequency range. The detector's ability to distinguish the target talker from interfering talkers could possibly be improved by modeling it instead on a dynamic fundamental frequency range, and by adding additional talker specific features such as voice timbre, glottal waveform, and habitual speech patterns.

Table 2: Average SRT in dB for hearing impaired (HI) and normal hearing (NH) subjects by processing method.

key	processing method	HI	NH
—●—	unprocessed	-21.6	-14.6
—●—	coherent	-18.1	-15.6
—●—	coherent in quiet	-24.5	-17.8
—●—	noise reduction	-20.6	n/a

6. REFERENCES

[1] J. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP J. Applied Sig. Proc.*, vol. 18, pp. 2915–2929, 2005.

[2] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–2905, 1997.

[3] S. M. Schimmel, K. R. Fitz, and L. E. Atlas, "Frequency reassignment for coherent modulation filtering," in *ICASSP*, 2006, vol. 5, pp. 261–264.

[4] S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *ICASSP*, 2007, vol. 4, pp. 605–608.

[5] S. M. Schimmel and L. E. Atlas, "Coherent envelope detection for modulation filtering of speech," in *ICASSP*, 2005, vol. I, pp. 221–224.

[6] S. M. Schimmel, *Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices*, Ph.D. thesis, University of Washington, Seattle, 2007.

[7] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Am.*, vol. 43, no. 4, pp. 829–834, 1968.

[8] M. Abe and J. O. Smith, III, "Design criteria for the quadratically interpolated FFT method (I): Bias due to interpolation," Tech. Rep. STAN-M-114, Stanford, CCRMA, 2004.

[9] B. G. Quinn, "Estimating frequency by interpolation using Fourier coefficients," *IEEE Trans. Sig. Proc.*, vol. 42, no. 5, pp. 1264–1268, 1994.

[10] J. Roberts, "Dynamic tracking filter as a low-threshold demodulator in F.M. F.D.M. satellite systems," *Proceedings of the IEE*, vol. 115, pp. 1597–1606, 1968.

[11] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech and Audio Proc.*, vol. 8, no. 3, pp. 240–254, 2000.

[12] J. A. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoustics, Speech, and Sig. Proc.*, vol. 22, pp. 330–338, 1974.

[13] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoustics, Speech, and Sig. Proc.*, vol. 26, pp. 354–358, 1978.

[14] D. Veeneman and B. Mazor, "A fully adaptive comb filter for enhancing block-coded speech," *IEEE Trans. Acoustics, Speech, and Sig. Proc.*, vol. 37, no. 6, pp. 955–957, 1989.

[15] W. R. Drennan, J. H. Won, V. K. Dasika, and J. T. Rubinstein, "Effects of temporal fine structure on the lateralization of speech and on speech understanding in noise," *Journal of the Association for Research in Otolaryngology*, 2007.

[16] C. W. Turner, B. J. Gantz, C. Vidal, A. Behrens, and B. A. Henry, "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1729–1735, 2004.

[17] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 467–477, 1971.