JSLHR

Article

Tackling the Combined Effects of Reverberation and Masking Noise Using Ideal Channel Selection

Oldooz Hazrati[a] and Philipos C. Loizou[a]

Purpose: In this article, a new signal-processing algorithm is proposed and evaluated for the suppression of the combined effects of reverberation and noise.

Method: The proposed algorithm decomposes, on a short-term basis (every 20 ms), the reverberant stimuli into a number of channels and retains only a subset of the channels satisfying a signal-to-reverberant ratio (SRR) criterion. The construction of this criterion assumes access to a priori knowledge of the target (anechoic) signal, and the aim of this study was to assess the full potential of the proposed channel-selection algorithm, assuming that this criterion could be estimated accurately. Listening tests with normal-hearing listeners were conducted to assess the performance of the proposed algorithm in highly reverberant conditions (T_{60} = 1.0 s), which included additive noise at 0 and 5 dB signal-to-noise ratios (SNRs).

Results: A substantial gain in intelligibility was obtained in both reverberant and combined reverberant and noise conditions. The mean intelligibility scores improved by 44 and 33 percentage points at 0 and 5 dB SNR reverberation + noise conditions. Feature analysis of the consonant confusion matrices revealed that the transmission of voicing information was most negatively affected, followed by manner and place of articulation.

Conclusions: The proposed algorithm produced substantial gains in intelligibility, and this benefit was attributed to the ability of the proposed SRR criterion to detect accurately voiced or unvoiced boundaries. It was postulated that detection of those boundaries is critical for better perception of voicing information and manner of articulation.

Key Words: reverberation, noise, dereverberation algorithms

Reverberation is present in most daily listening situations. Reverberation can cause significant changes in speech quality and can have a very negative impact on speech intelligibility as it blurs, for instance, temporal and spectral cues and flattens formant transitions (Nabelek, Letowski, & Tucker, 1989). Although moderate amounts of reverberation do not affect speech recognition performance by normal-hearing listeners, reverberation has a detrimental effect on speech intelligibility by listeners with hearing impairment and elderly listeners (Assmann & Summerfield, 2004; Nabelek, 1993) as well as by automatic speech recognition systems (Palomäki, Brown, & Parker, 2004). The negative effects of reverberation on intelligibility vary across age (Neuman & Hochberg, 1983; Neuman, Wroblewski, Hajicek, & Rubinstein, 2010) and between native and nonnative listeners (Nabelek & Donahue, 1984).

Nabelek and Letowski (1985) studied the effects of reverberation on vowel recognition by 10 elderly adults with binaural sensorineural hearing loss and found that the mean vowel recognition score obtained in a reverberation time (T_{60}) of 1.2 s was approximately 12 percentage points lower than the mean score obtained in the nonreverberant (anechoic) conditions. Compared with vowels, consonants are generally more affected by reverberation. The stop consonants, for instance, are more susceptible to reverberation distortion than other consonants, particularly in syllable-final position. This is because reverberation “fills in” the gaps present during stop closures. When noise is added to reverberation, listeners make different consonant confusions from those made in reverberation or in noise (Nabelek et al., 1989). That is, noise generally masks speech differently than reverberation. The combined effects of reverberation and noise are

[a]The University of Texas at Dallas, Richardson

Correspondence to Philipos C. Loizou: loizou@utdallas.edu

Editor: Sid Bacon
Associate Editor: Eric Healy

Received March 25, 2011
Accepted August 15, 2011
DOI: 10.1044/1092-4388(2011/11-0073)

quite detrimental to intelligibility (Nabelek & Mason, 1981; Nabelek & Pickett, 1974a). In a recent study with normal-hearing children and adults, Neuman et al. (2010) assessed speech intelligibility in reverberation + noise conditions in terms of speech reception threshold (SRT). When comparing the results to SRT norms obtained by adults in anechoic conditions, Neuman et al. (2010) reported a signal-to-noise ratio (SNR) loss[1] of 1.5–3 dB for adults and 7.5–9.5 dB for young children (age 6 years) when reverberation (T_{60} = 0.3–0.8 s) was added. The SNR loss decreased as a function of age.

Addressing the degradation in speech intelligibility and quality due to reverberation has given rise to several dereverberation algorithms (e.g., see Benesty, Sondhi, & Huang, 2007; Jin & Wang, 2009; Kollmeier & Koch, 1994; Naylor & Gaubitch, 2010). Dereverberation by means of *inverse filtering*—or passing a reverberant signal through a finite impulse response (FIR) filter that inverts the reverberation process—remains one of the most commonly used methods (Miyoshi & Kaneda, 1988). However, the main drawback of inverse filtering methods is that the acoustic impulse response must be known in advance or, alternatively, needs to be "blindly" estimated. Such algorithms, however, have severe limitations because room impulse responses (RIRs), particularly in highly reverberant rooms, have thousands of filter taps, making their inversion a computationally expensive task. Furthermore, some RIRs exhibit nonminimum phase characteristics. Thus, techniques that do not rely on inversion of the RIR are more attractive and more practical.

An alternative technique based on channel selection is explored in this article. Such a technique is attractive because it does not rely on the inversion of the RIR. The proposed method is based on decomposing, in short time segments (every 20 ms), the reverberant signal into a number of channels (via a fast Fourier transform [FFT]) and retaining only a subset of channels at each segment. The proposed criterion for selecting the appropriate channels is based on instantaneous measurements of the signal-to-reverberant ratio (SRR). Envelopes (computed using the FFT magnitude spectrum of each 20-ms segment) corresponding to channels with SRR larger than a preset threshold are selected, whereas envelopes corresponding to channels with SRR smaller than the threshold are zeroed out. The SRR reflects the ratio of the energies of the signal originating from the early (and direct) reflections and the signal originating from the late reflections. Note that the resulting reverberant signal is composed of the superposition of these two aforementioned signals. Hence, the underlying motivation in using the proposed SRR criterion is to retain the signal components arising from the early reflections while discarding the signal components generated from late reflections. Early reflections are known to benefit speech intelligibility in binaural hearing (e.g., see Litovsky, Colburn, Yost, & Guzman, 1999) for normal-hearing listeners, whereas late reflections are known to be detrimental to speech intelligibility as they are responsible predominantly for smearing the temporal envelopes and filling the gaps (e.g., closures) in unvoiced segments (e.g., stops) of the utterance.

The above SRR criterion for channel selection was used and evaluated in our prior study (Kokkinakis, Hazrati, & Loizou, 2011) with listeners who have cochlear implants. That study, however, only evaluated the effects of reverberation—that is, with no additive noise present. In the present study, we evaluated the proposed channel-selection criterion using normal-hearing listeners in conditions wherein additive noise as well as reverberation are present. Nonsense syllables were used for testing to avoid ceiling effects. The aim of this study was twofold: (a) to determine the effectiveness of the proposed channel-selection criterion in suppressing or minimizing the combined effects of reverberation and noise and (b) to determine which consonant feature (voicing, manner of articulation, or place of articulation) is affected the most in reverberation + noise conditions.

Method

Listeners

Eight normal-hearing listeners with pure-tone thresholds less than 25 dB HL (at frequencies of 250 Hz up to 8 kHz), all native speakers of American English, were recruited for the intelligibility tests. Their ages ranged from 18 to 26 years, and all subjects were paid for their participation. The majority of the subjects were undergraduate students from The University of Texas at Dallas.

Stimuli

Telephone band-limited (300–3400 Hz) syllables in /aCa/ context were used for testing. The consonant set included 16 consonants recorded in /aCa/ context, where C = /p, t, k, b, d, g, m, n, dh, l, f, v, s, z, sh, jh/. All consonants were produced by an American male speaker and were recorded in a soundproof booth using Tucker-Davis Technologies recording equipment. The consonants were originally sampled at 25 kHz and down-sampled to 8 kHz. To simulate the receiving frequency characteristics of telephone handsets, all clean and corrupted signals were filtered by the modified intermediate reference system (IRS) filters (ITU-T Recommendation P.48, 1996). Telephone

[1] *SNR loss* reflects the increase in SNR required to attain 50% correct performance due to reverberation. This is relative to the SNR required by normal-hearing adults in anechoic, noise-alone conditions (SRT).

band–limited consonants were used to avoid ceiling effects.

The reverberant signals were generated by convolving the clean signals with real RIRs, recorded by Van den Bogaert, Doclo, Wouters, and Moonen (2009), with average reverberation time of T_{60} = 1.0 s and direct-to-reverberant (DRR) ratio of –0.49 dB for a 5.50 m × 4.50 m × 3.10 m (length × width × height) room. The distance between the single-source signal and the microphone was 1 m. Speech-shaped noise was added to the reverberant signals at 0 dB SNR and 5 dB SNR—that is, the reverberant speech signal served as the target signal in the SNR computation.

Proposed Algorithm Based on Channel Selection

The proposed algorithm, henceforth called the *ideal channel-selection (ICS) algorithm,* is depicted in Figure 1. It is termed *ideal* to indicate that a priori information about the target signal is used. First, the clean and corrupted signals are segmented into 20-ms frames (with 50% overlap between frames) using a Hanning window, and a discrete Fourier transform (DFT) is computed. Note that in the frequency domain, the DFT decomposes the signal into N frequency bins, or channels (alternatively, an M-channel filter bank could be used in place of the DFT), where N is the duration of the frame in samples. Of the $N/2$ available channels (due to the DFT symmetry), a subset is selected based on the SRR criterion,[2] which is computed as follows:

$$SRR(f,t) = 10\log_{10}\frac{|S(f,t)|^2}{|R(f,t)|^2}, \tag{1}$$

where t indicates the frame index, f indicates the channel or frequency bin index, and $S(f,t)$ and $R(f,t)$ denote the complex DFT spectra of the clean (anechoic) and corrupted (reverberant or reverberation + noise) signals, respectively. Envelopes (computed using the FFT magnitude spectrum of each 20-ms segment) corresponding to channels with SRR > T are selected, whereas envelopes corresponding to channels with SRR ≤ T are discarded, where T denotes a preset threshold value. Mathematically, this can be expressed by applying a gating or binary gain (BG) function to the spectrum of the reverberant signal as follows:

$$\widehat{S}(f,t) = R(f,t).BG(f,t), \tag{2}$$

where

$$BG(f,t) = \begin{cases} 1, & if\ SRR(f,t) > T \\ 0, & otherwise \end{cases} \tag{3}$$

and where T represents the threshold value expressed in dB. To reconstruct the enhanced (dereverberated) signal in the time domain, the inverse DFT of $\widehat{\mathrm{S}}(\mathrm{f},\mathrm{t})$ is computed, and the signal is finally synthesized using the overlap-add (OLA) method (MATLAB implementation of the above ICS algorithm is available from our website[3]). It is important to stress that the above binary time-frequency gain function (Equation 3) is applied to the reverberant spectrum and does not explicitly “clean out” reverberation but, rather, selects the reverberant channels that satisfy the SRR criterion (Equation 3).

The above operation of (ideal) binary gating is also known in the literature as the *ideal binary mask,* or *ideal time-frequency mask,* and it has been used extensively in computational models of auditory scene analysis (see review in Wang & Brown, 2006). The ideal binary mask uses as a channel-selection criterion the instantaneous SNR computed at each time-frequency unit and has been used in applications in which the objective is to segregate a target speaker from a mixture (see Brungart, Chang, Simpson, & Wang, 2006; Li & Loizou, 2008a). The SNR criterion is clearly not appropriate for the reverberation-alone conditions, considering there are no additive maskers present. Given these differences, we refer to our algorithm as the *ideal channel-selection algorithm* rather than as the ideal binary mask algorithm. Both algorithms select in each frame a subset of channels, but the selection is made using different criteria.

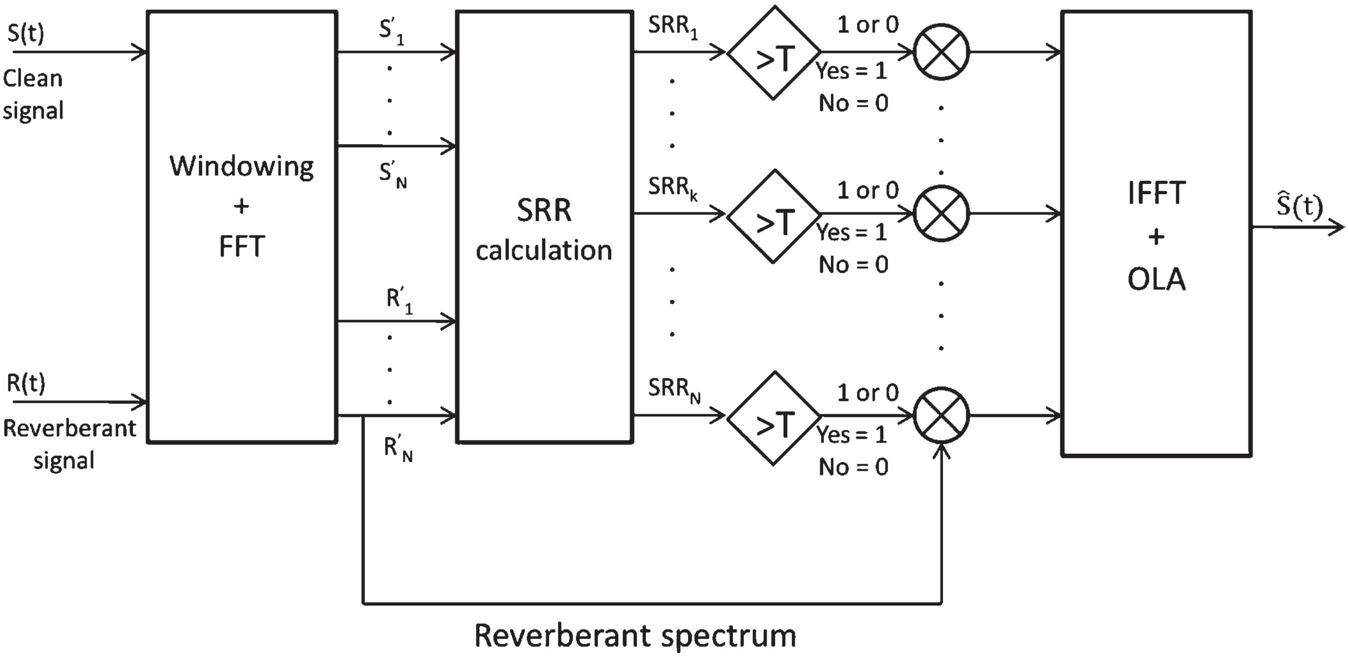
The choice of the threshold T in Equation 3 is important in the construction and application of the proposed channel-selection criterion. To illustrate this, we show in the Figure 2 example synthesized waveforms of the syllable /a p a/, with the threshold set to T = –8 dB (Panel [d]) and T = 0 dB (Panel [e]). As shown, the latter threshold (0 dB) is aggressive, considering that apart from discarding the corrupted unvoiced segments and associated gaps, it also eliminates speech in the voiced frames, which in turn leads to distortion of the processed signal. In contrast, the use of T = –8 dB seems to eliminate the overlap-masking effects caused by the overlapping of succeeding segments of speech (in our case, the stop /p/) by the preceding phonetic segments (vowel /a/ in this example). As shown in Figure 2, by appropriately thresholding the SRR function (shown in Panel [c]), we can reliably identify the vowel/consonant boundaries even in highly reverberant settings (T_{60} = 1.0 s).

Figure 3 shows example synthesized waveforms of the syllable /a s a/ corrupted by reverberation and additive

[2] Note that there are two major differences between our definition of SRR and the conventional SRR definition (Naylor & Gaubitch, 2010, Chapter 2). First, we do not use the direct-path signal, and second, the SRR given in this article is defined in the frequency domain for each T-F unit and is computed for each frame of the stimulus data.

[3] www.utdallas.edu/~loizou/cimplants/

Figure 1. Block diagram of the proposed ideal channel-selection (ICS) algorithm.



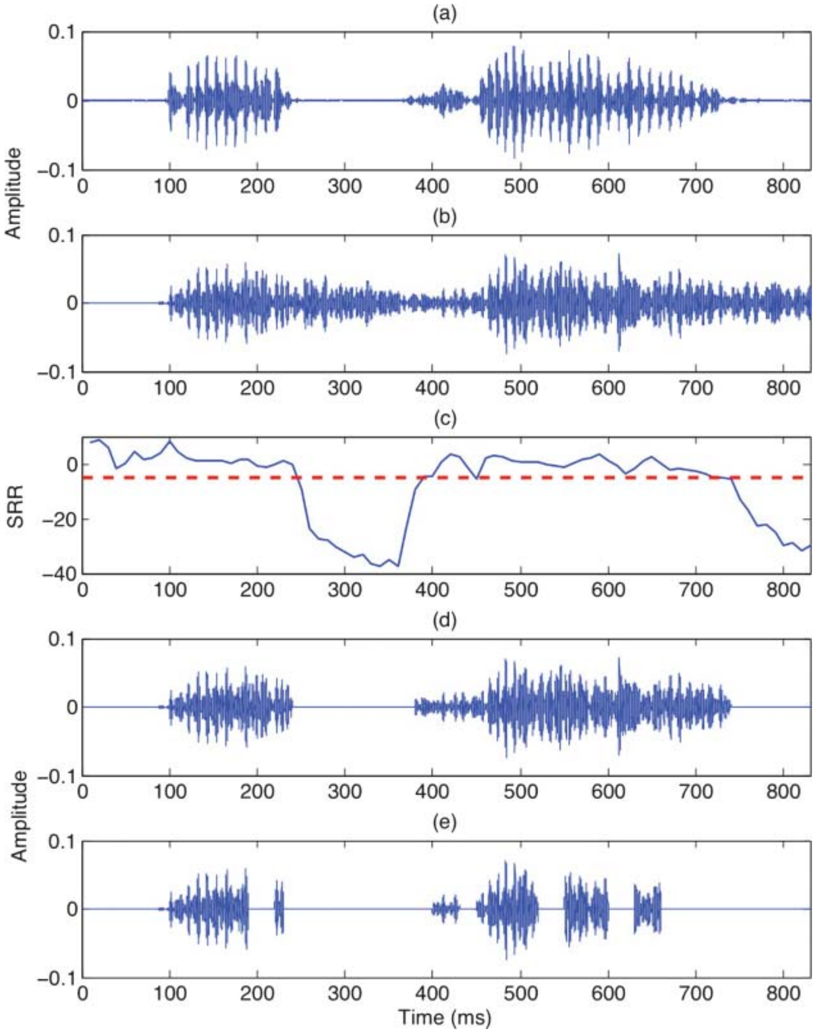
noise at +5 dB SNR. Waveforms from a low-frequency channel (f = 500 Hz) are shown in the left column, and waveforms from a high-frequency channel (f = 3060 Hz) are shown in the right column. As can be seen in Panels (d) and (h), the retained (by the channel-selection process) waveforms are still corrupted by noise; however, the vowel/consonant boundaries are preserved. The gap in the /s/ spectrum, for instance, at t = 300–500 ms is maintained in the retained waveform (compare Panels [a] and [d]).

The motivation for choosing the SRR criterion to guide the channel-selection process is as follows. The SRR provides, approximately, a simple measure of the ratio of the signal energy conveyed by the early reflections (and the direct sound) to that contained in the late reflections. It seems reasonable, then, to select a given channel only when the signal energy produced by early reflections dominates the energy produced by late reflections originating from the preceding signal. This is demonstrated in Figures 2 and 3. In Figure 2, for instance, during the /p/ closure (t = 245–362 ms), the reverberant signal contains a significant amount of energy caused by leakage from the preceding vowel (overlap masking). This energy is introduced primarily by the late reflections. The SRR takes extremely low values (–10 to –40 dB) during that period of the /p/ closure, wherein the contributions from the late reflections dominate. Consequently, by discarding a channel when the SRR is extremely low, we are reducing (and, to some extent, minimizing) the overlap-masking effects. In contrast, a large SRR value suggests dominance of the energy from the direct signal (and early reflections), as is often the case during the voiced segments (e.g., vowels) of the utterance (overlap masking may occur during voiced segments due to the energy originating from the preceding consonant; however, its effect is minimal). Consequently, channels containing energy from early reflections are retained (see, for instance, the vowel segment from t = 450 ms to t = 738 ms).

In reality, the denominator in the SRR contains energy from both early and late reflections, but nonetheless, we assume that the contribution of the early reflections is small. This assumption holds for the most part during unvoiced phonetic segments containing spectral gaps (e.g., stop closures), particularly in the low frequencies where the vowel formants reside and overlap-masking effects dominate. Ideally, it would be desirable to decouple the contributions of the early and late reflections, but that is not straightforward or easy to do, particularly when the reverberation time (T_{60}) is long. For that reason, the entire reverberant signal is used in the denominator of the SRR for practical purposes. Similarly, we assume that the energy produced from the early reflections is close to that produced by the direct path. The proposed experiments will test whether the above approximations and assumptions hold.

A number of alternative criteria to the SRR criterion have been proposed and evaluated by Mandel, Bressler,

Figure 2. Band-pass filtered waveforms (f = 1 kHz, bandwidth = 138 Hz) of /a p a/ for (a) clean, (b) reverberant (T_{60} = 1.0 s), (d) reverberant signal processed by ideal channel-selection (ICS) with T = –8 dB, and (e) reverberant signal processed by ICS with T = 0 dB. Panel (c) shows the instantaneous signal-to-reverberant ratio (SRR) values (horizontal line indicates the fixed threshold set at –8 dB).

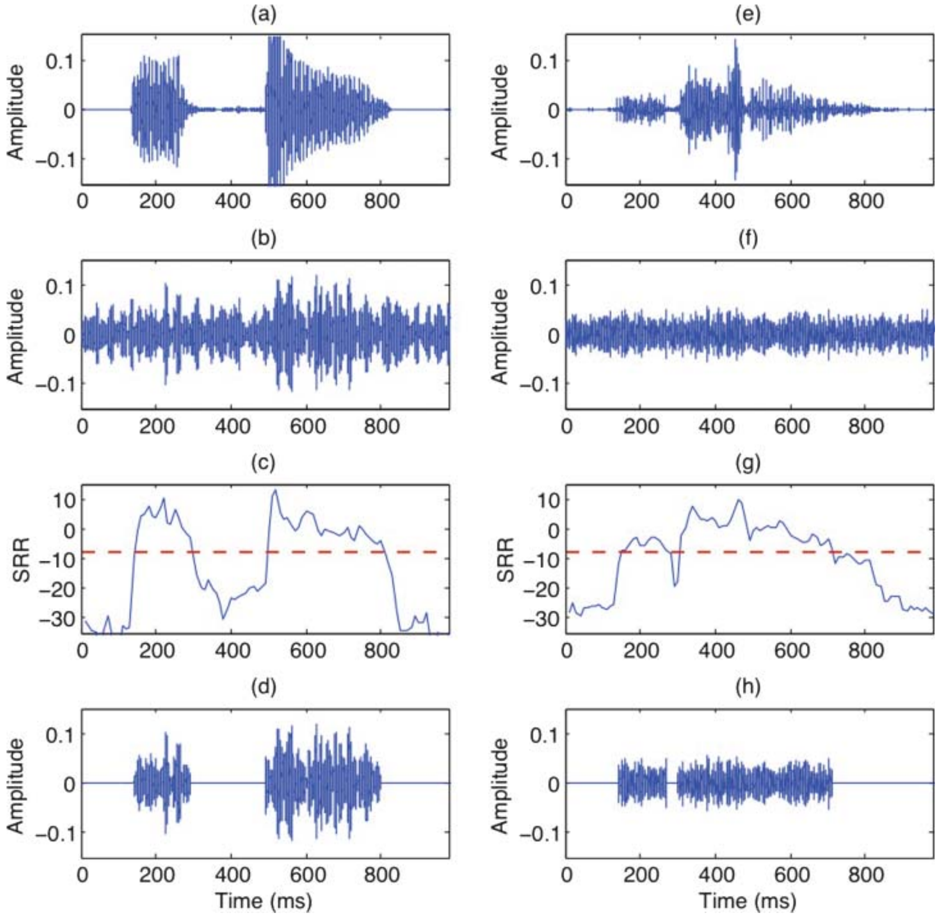


Shinn-Cunningham, and Ellis (2010). However, these criteria were evaluated in the context of improving the performance of automatic speech recognition systems rather than as a means for improving speech intelligibility. In addition, most of the criteria proposed by Mandel et al. (2010) required access to the RIR. Hence, estimating such criteria poses great challenges. In contrast, the construction of the SRR criterion does not require access to the RIR.

As mentioned earlier, we denote the proposed algorithm as the ideal channel-selection (ICS) algorithm, where the term *ideal* is used to indicate that a priori knowledge is used to construct the SRR. In practice, the SRR needs to be estimated from the reverberant signal alone. The aim of the proposed experiments is to assess the full potential of the SRR criterion in terms of intelligibility benefit. If a large benefit is observed, that would suggest that significant efforts need to be devoted to developing techniques for accurately estimating the SRR. The resulting data from the proposed experiments will provide the upper bound in performance that can be obtained when the SRR criterion is estimated accurately.

Procedure. The listeners participated in 15 conditions, which included clean anechoic stimuli, reverberant stimuli, stimuli corrupted by noise alone (at 2 SNRs), and reverberation + noise (at 2 SNRs) stimuli. Six additional conditions involved ICS-processed stimuli (reverberant and reverberant + noise) using two different threshold values (T = –8 dB and 0 dB). We denote the reverberant stimuli as R, the stimuli corrupted by noise alone as N,

Figure 3. Panels in the left column show band-pass filtered waveforms of /a s a/ in a low-frequency channel (f = 500 Hz, bandwidth = 81 Hz), and panels in the right column show band-pass filtered waveforms in a high-frequency channel (f = 3060 Hz, bandwidth = 418 Hz). Panels (a) and (e) show the clean waveforms, Panels (b) and (f) show the reverberation + noise (1.0 s, 5 dB) waveforms, and Panels (d) and (h) show the reverberation + noise (1.0 s, 5 dB) waveforms processed by ICS with T = −8 dB. Panels (c) and (g) show the instantaneous SRR values (horizontal line indicates the fixed threshold set at −8 dB).



and the reverberation + noise stimuli as R + N. Three other conditions were included for comparative purposes based on a commonly used spectral subtractive (Wu & Wang, 2006) algorithm for suppressing reverberation. This algorithm was applied to the reverberant and R + N stimuli. The spectral subtraction algorithm has been found to be effective in removing the impact of late reverberation (Wu & Wang, 2006) and was used in this study as an additional control condition.[4]

The consonants were presented to the listeners in random order. Six repetitions per condition were used. The presentation order of the various conditions was randomized across subjects. A practice session, in which the clean (anechoic) consonants were presented to the listeners, preceded the actual test. To collect responses, a graphical user interface (GUI) was used that allowed the subjects to identify the consonants they heard by clicking on the corresponding button on the GUI. All listening experiments were performed in a soundproof room (Acoustic Systems, Inc.) using a PC connected to a Tucker-Davis System 3. Stimuli were presented to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. The test session lasted for approximately 2 hr. A short break was given to the subjects every 30 min to minimize listener fatigue.

[4] A two-stage algorithm was originally proposed in Wu and Wang (2006). In the first stage, an inverse filtering algorithm was adopted for reducing coloration effects followed by a spectral-subtractive algorithm in the second stage for reducing late-reverberation effects. Note that the second stage was designed to subtract out the late reflections from the reverberated signal rather than subtract out additive noise. We were not able to obtain satisfactory performance via the inverse-filtering stage due to the long impulse response used in our study corresponding to a long reverberation time (T_{60} = 1.0 s). For that reason, we implemented only the second stage.

Results and Discussion

Consonant Identification in Reverberation and Noise

The results, expressed in terms of the mean percentage of consonants identified correctly, are shown in Figure 4. The bar labeled “clean” represents the mean score obtained in anechoic conditions. A two-way, repeated measures analysis of variance (ANOVA)[5] indicated significant effect of SNR, $F(2, 14) = 63.2$, $p < .0005$; a significant effect of ICS threshold, $F(2, 14) = 257.1$, $p < .0005$; and a significant interaction, $F(4, 28) = 44.8$, $p < .0005$. As shown in Figure 4, the ICS algorithm improved speech intelligibility in all conditions, including the reverberation-alone (i.e., with no additive noise) condition. The choice of ICS threshold affected performance in the two noisy conditions but had little effect in quiet conditions because of ceiling effects. For that reason, an interaction was observed between the SNR level and ICS threshold.

ICS improved intelligibility in all conditions tested. Post hoc tests, according to Tukey’s honestly significant difference, were conducted to assess the differences in scores between conditions. The intelligibility scores obtained in the reverberation-alone condition improved by more than 7 percentage points when the ICS algorithm was used. This difference was found to be statistically significant ($p = .003$). Larger improvements with the ICS algorithm were noted in the R + N conditions. Recognition scores improved from 61.7% correct to 94.5% correct at 5 dB SNR, and from 51.4% to 95.2% at 0 dB SNR. In all cases, the intelligibility improvement (relative to that with unprocessed stimuli) by the proposed ICS algorithm was found to be statistically significant ($p < .005$) when implemented with either threshold value ($T = –8$ dB or 0 dB). In the 0-dB R + N condition, the score obtained with the ICS threshold set to $T = –8$ dB was found to be significantly ($p < .0005$) higher than the score obtained with the ICS threshold set to $T = 0$ dB. In the 5-dB R + N condition, the score obtained with the ICS threshold set to $T = –8$ dB was not found to be significantly ($p > .05$) higher than the score obtained with the ICS threshold set to $T = 0$ dB. High performance was consistently obtained across all conditions tested when the ICS threshold was set to $T = –8$ dB. In all conditions tested, performance with the ICS was near that obtained by listeners in anechoic conditions—that is, near 96% correct.

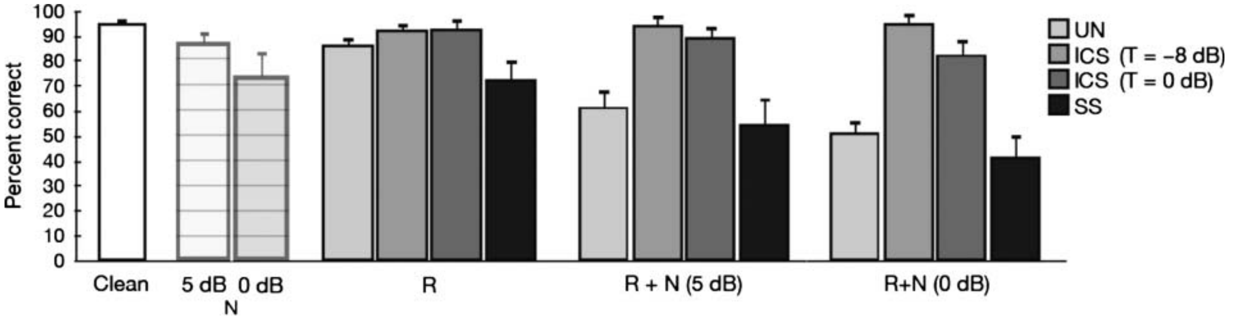
Figure 4 also shows performance obtained in the condition wherein the stimuli were corrupted only by noise (no reverberation). The intelligibility scores were 86.59% and 73.05% at SNRs of 5 dB and 0 dB, respectively. These scores decreased to 62% and 51%, respectively, after adding reverberation. In both noise conditions (0 dB SNR and 5 dB SNR), scores were reduced by 30% after reverberation was added. Noise and reverberation degrade intelligibility in a complementary fashion. That is, regions in the spectrum that were not originally corrupted by reverberation are affected or masked by noise, leading to a severe degradation in intelligibility (30% reduction in our study). As reported by others, the combined effects of noise and reverberation are greater than the sum of both effects taken separately (Nabelek & Mason, 1981; Nabelek & Pickett, 1974b). This was also confirmed with the data in our study. Table 1 shows the effects of reverberation, noise, and combined effects of reverberation and noise for individual subjects. These effects were computed by assessing the decrement in performance relative to the performance obtained in anechoic conditions. The combined effect was computed, for instance, as the difference between the scores obtained in the R + N condition and the scores in the anechoic condition. For nearly all subjects (except S6) and for both SNRs tested, the combined effects were greater than the sum of the reverberation and noise effects.

As can be seen from Figure 4, the performance of the spectral-subtractive dereverberation algorithm (Wu & Wang, 2006) was not satisfactory. The scores obtained using the spectral-subtraction algorithm in the reverberation-alone condition were significantly ($p < .005$) lower than the scores obtained using the unprocessed reverberant stimuli. This is partly because applying spectral subtraction may introduce signal distortion and may, therefore, produce a drop in the consonant identification scores. Second, the reverberant conditions examined in this study were quite challenging (this algorithm was originally tested in shorter reverberation times [$T_{60} = 0.2–0.4$ s] by Wu & Wang [2006]). Performance in R + N conditions was even less satisfactory. We believe that it was because the SS algorithm was not originally developed to cope with low SNR conditions, as it had been originally tested at a high SNR (20 dB).

In brief, the proposed ICS algorithm was found to produce substantial gains in intelligibility in both reverberation-alone conditions and in conditions involving additive noise (see Figure 4). This outcome was consistent with the benefit observed by listeners with cochlear implants in our prior study (Kokkinakis et al., 2011). We attribute the intelligibility benefit to the ability of the SRR criterion to detect accurately voiced/unvoiced boundaries (see Figure 2). In continuous speech, reliable access to these vowel/consonant boundaries has been found to be critical for lexical segmentation and word retrieval (Li & Loizou, 2008b; Stevens, 2002).

[5] We also ran the statistics using arcsine-transformed scores, but the results and conclusions remained the same.

Figure 4. Mean intelligibility scores obtained in the various conditions involving reverberation (R), noise (N), and combined reverberation and noise (R + N). The leftmost bar shows performance obtained in (clean) anechoic conditions. Scores obtained with unprocessed stimuli are labeled as UN, and scores obtained with stimuli processed via the spectral-subtractive algorithm are labeled as SS.



Analysis of Consonant Errors

The consonant confusion matrices were analyzed in terms of percentage of transmitted information as per Miller and Nicely (1955), and the mean feature scores for place of articulation, manner of articulation, and voicing features are presented in Figure 5. A two-way repeated measures ANOVA indicated a significant effect of SNR, $F(1, 7) = 8.6$, $p = .022$; a significant effect of feature error, $F(2, 14) = 13.3$, $p = .001$; and a nonsignificant interaction, $F(2, 14) = 0.649$, $p = .538$. As can be seen from Figure 5, all three features, especially the voicing feature, were significantly affected, $F(2, 14) = 13.3$, $p = .001$, in the R + N conditions. In the presence of reverberation, place-of-articulation scores were generally higher than the manner and voicing scores.

Overall, on the basis of Figure 5, we can conclude that the transmission of voicing information is mostly affected in the R + N conditions. We attribute this to overlap-masking effects, which are largely responsible for filling the gaps (e.g., stop closures) that are present in some consonants (e.g., stops), making it difficult to distinguish between, for instance, the unvoiced stops (e.g., /t/) and the voiced stops (e.g., /d/). The filled gaps (by reverberation and noise) clearly affect the perception of voice onset time, and it is well known that in intervocalic stops, the duration of voice onset time as well as the duration of aspiration are effective cues signaling voicing contrast (Borden, Harris, & Raphael, 1994). These cues are severely corrupted by the combined effects of reverberation and noise.

All feature errors in place, manner, and voicing were compensated by the use of the ICS algorithm. An average improvement of about 8 percentage points in amount of transmitted information was achieved in the reverberation-alone condition. A larger improvement was noted at 0 dB SNR and 5 dB SNR, respectively, and that amounted to approximately 64 and 43 percentage points (respectively) for place, 43 and 36 points (respectively) for manner, and 65 and 54 points (respectively) for voicing in R + N conditions (on average, only 4%, 2%, and 6% below the scores obtained for place, manner, and voicing features, respectively, in the anechoic quiet conditions).

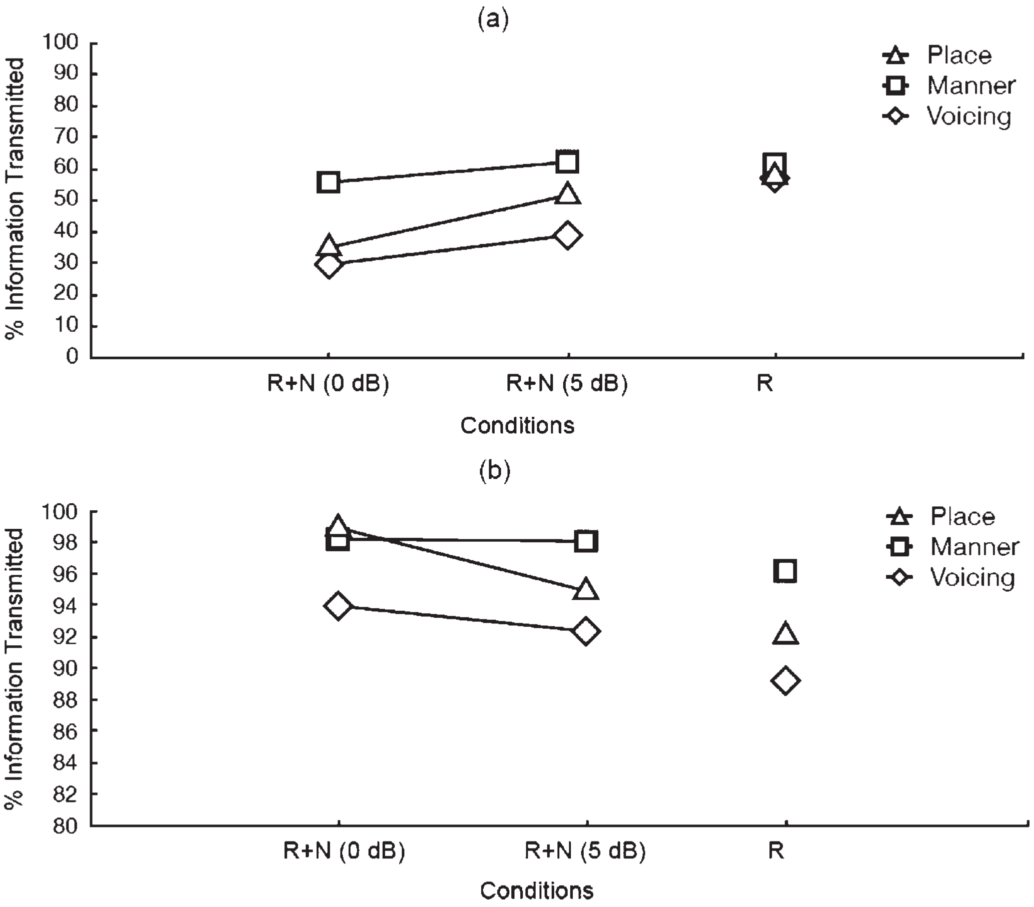
Table 1. Effects of reverberation (R), noise (N), and the combination of reverberation and noise (R + N) on consonant identification (%) for individual subjects. Low scores reflect small effects relative to the scores obtained in the anechoic condition.

Subject	R	N (5 dB)	R + N (5 dB)	N (0 dB)	R + N (0 dB)
1	4.2	12.5	43.8	15.6	47.9
2	2.1	5.2	32.3	14.6	41.7
3	7.3	0.0	32.3	17.7	42.7
4	15.6	8.3	40.6	31.3	54.2
5	7.3	10.4	33.3	27.1	43.8
6	9.4	12.5	31.3	34.4	36.5
7	10.4	11.5	31.3	27.1	39.6
8	14.6	9.4	24.0	10.4	44.8

Predicting the Intelligibility of R + N Speech

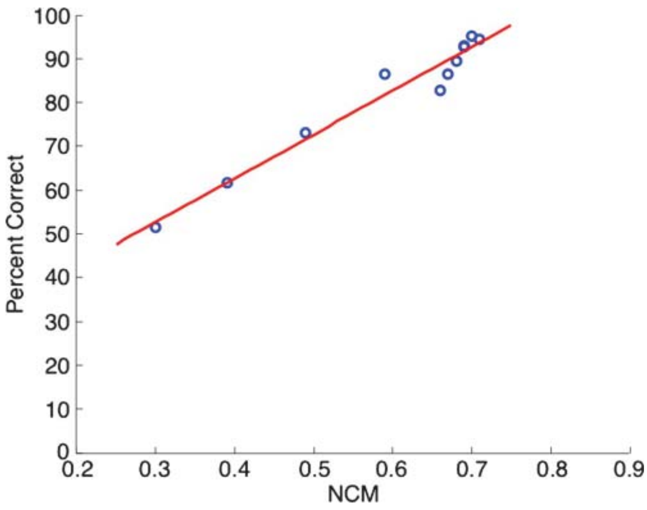
The speech-transmission index (STI) has been shown in a number of studies to predict reliably the intelligibility of speech in reverberation, in noise, or in both (Houtgast & Steeneken, 1985). Aside from the STI, not many intelligibility measures exist that can predict the combined effects of reverberation and noise. In this study, we evaluated the performance of a speech-based STI that has been found previously to correlate highly with the intelligibility of noise-masked and noise-suppressed speech (Chen & Loizou, 2010; Ma, Hu, & Loizou, 2009). More precisely, we selected the normalized covariance measure (NCM), which is similar to the STI in that it computes the STI as

Figure 5. Relative information transmitted for (a) unprocessed stimuli in reverberant (R) and combined reverberation + noise conditions (R + N) and (b) stimuli processed via the ICS algorithm in different conditions (for better clarity, the *y*-axis range is limited within 80% and 100%).



a weighted sum of transmission index values determined from the envelopes of the probe and response signals in each frequency band (Goldsworthy & Greenberg, 2004). Unlike the traditional STI, however, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function, the NCM is based on the covariance between the probe (input) and response (output) envelope signals. The NCM has not been evaluated previously in situations where noise is present along with reverberation. Figure 6 shows the scatter plot of the mean consonant intelligibility scores obtained in all conditions (except the control conditions) against the corresponding mean NCM values. A linear fit is shown, but alternatively, a sigmoidal-shaped function could be used to fit the data (the computed correlation coefficient was found to be the same with either fitting function). The resulting Pearson's correlation coefficient was found to be quite high, $r = .98$. These data clearly show that the NCM is an effective measure for predicting not only speech intelligibility in noise (Ma et al., 2009) but also

Figure 6. Scatter plot of the mean consonant recognition scores obtained in this study in the various conditions against the corresponding normalized covariance measure (NCM) values.



speech intelligibility corrupted by both reverberation and noise.

Conclusions

The combined effects of reverberation and noise have been found in this study to be quite detrimental to consonant recognition, an outcome consistent with prior studies (Nabelek & Mason, 1981; Nabelek & Pickett, 1974b). A signal-processing algorithm was proposed for the suppression of combined reverberation and noise. This algorithm is based on the decomposition of the reverberant stimuli into a number of frequency channels and the selection of channels with SRR exceeding a preset threshold (–8 dB). Channels with SRR values falling below the threshold were discarded. Hence, in the proposed algorithm, neither reverberation nor noise was explicitly suppressed or attenuated in any way because the channel-selection process was applied directly to the R + N stimuli. When presented to normal-hearing listeners, the synthesized stimuli have been found to yield substantial gains in consonant identification. This outcome suggests that the combined effects of reverberation and noise do not completely mask important speech information. The channel-selection process, as guided by the SRR criterion, is a powerful process that can uncover quite effectively important speech information from the corrupted (by reverberation and noise) stimuli. Analysis of the consonant confusion errors indicated that the proposed algorithm significantly improved the transmission of voicing information, along with manner and place of articulation. Much of the intelligibility benefit was attributed to the ability of the SRR channel-selection criterion to accurately detect and preserve voiced/unvoiced boundaries, often smeared in the presence of reverberation.

Acknowledgment

This work was supported by National Institute of Deafness and Other Communication Disorders Grant R01 DC 010494, awarded to the second author.

References

Assmann, P. F., & Summerfield, Q. (2004). The perception of speech under adverse acoustic conditions. In S. Greenberg, W. A. Ainsworth, A. N. Popper, & R. R. Fay (Eds.), *Speech processing in the auditory system* (pp. 231–308). New York, NY: Springer.

Benesty, J., Sondhi, M., & Huang, Y. (2007). *Handbook of speech processing*. Berlin, Germany: Springer.

Borden, G. J., Harris, K. S., & Raphael, L. J. (1994). *Speech science primer: Physiology, acoustics, and perception of speech*. Baltimore, MD: Williams & Wilkins.

Brungart, D., Chang, P., Simpson, B., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America, 120,* 4007–4018.

Chen, F., & Loizou, P. C. (2010). Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech. *The Journal of the Acoustical Society of America, 128,* 3715–3723.

Goldsworthy, R., & Greenberg, J. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America, 116,* 3679–3689.

Houtgast, T., & Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America, 77,* 1069–1077.

ITU-T P.48. (1996). Specification for an intermediate reference system. *Blue Book, 5,* 81–86.

Jin, Z., & Wang, D. L. (2009). A supervised learning approach to monaural segregation of reverberant speech. *IEEE Transactions on Audio, Speech, and Language Processing, 17,* 625–638.

Kokkinakis, K., Hazrati, O., & Loizou, P. C. (2011). A channel-selection criterion for suppressing reverberation in cochlear implants. *The Journal of the Acoustical Society of America, 129,* 3221–3232.

Kollmeier, B., & Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *The Journal of the Acoustical Society of America, 95,* 1593–1602.

Li, N., & Loizou, P. C. (2008a). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America, 123,* 1673–1682.

Li, N., & Loizou, P. C. (2008b). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *The Journal of the Acoustical Society of America, 124,* 3947–3958.

Litovsky, R., Colburn, S., Yost, W., & Guzman, S. (1999). The precedence effect. *The Journal of the Acoustical Society of America, 106,* 1633–1654.

Ma, J., Hu, Y., & Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America, 125,* 3387–3405.

Mandel, M. I., Bressler, S., Shinn-Cunningham, B., & Ellis, D. P. W. (2010). Evaluating source separation algorithms with reverberant speech. *IEEE Transactions on Audio, Speech, and Language Processing, 18,* 1872–1883.

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27,* 338–352.

Miyoshi, M., & Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Transactions on Speech and Audio Processing, 36,* 145–152.

Nabelek, A. K. (1993). Communication in noisy and reverberant environments. In G. A. Stubebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 15–28). Needham Heights, MA: Allyn & Bacon.

Nabelek, A. K., & Donahue, A. M. (1984). Perception of consonants in reverberation by native and non-native listeners. *The Journal of the Acoustical Society of America, 75,* 632–634.

Nabelek, A. K., & Letowski, T. R. (1985). Vowel confusions of hearing-impaired listeners under reverberant and non-reverberant conditions. *Journal of Speech and Hearing Disorders, 50,* 126–131.

Nabelek, A. K., Letowski, T. R., & Tucker, F. M. (1989). Reverberant overlap- and self-masking in consonant identification. *The Journal of the Acoustical Society of America, 86,* 1259–1261.

Nabelek, A. K., & Mason, D. (1981). Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. *Journal of Speech and Hearing Research, 24,* 375–383.

Nabelek, A. K., & Pickett, J. M. (1974a). Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners. *Journal of Speech and Hearing Research, 17,* 724–739.

Nabelek, A. K., & Pickett, J. M. (1974b). Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation, and hearing aids. *The Journal of the Acoustical Society of America, 56,* 628–639.

Naylor, P. A., & Gaubitch, N. D. M. (Eds.). (2010). *Speech dereverberation*. London, England: Springer.

Neuman, A. C., & Hochberg, I. (1983). Children's perception of speech in reverberation and combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. *The Journal of the Acoustical Society of America, 73,* 2145–2149.

Neuman, A. C., Wroblewski, M., Hajicek, J., & Rubinstein, A. (2010). Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. *Ear and Hearing, 31,* 336–344.

Palomäki, K. J., Brown, G. J., & Parker, J. P. (2004). Techniques for handling convolutional distortion with "missing data" automatic speech recognition. *Speech Communication, 43,* 123–142.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111,* 1872–1891.

Van den Bogaert, T., Doclo, S., Wouters, J., & Moonen, M. (2009). Speech enhancement with multichannel Wiener filter techniques in multi-microphone binaural hearing aids. *The Journal of the Acoustical Society of America, 124,* 360–371.

Wang, D. L., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. New York, NY: Wiley.

Wu, M., & Wang, D. L. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing, 14,* 774–784.

Tackling the Combined Effects of Reverberation and Masking Noise Using Ideal Channel Selection

Oldooz Hazrati, and Philipos C. Loizou

J Speech Lang Hear Res 2012;55;500-510; originally published online Jan 9, 2012;
DOI: 10.1044/1092-4388(2011/11-0073)