Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices

Kostas Kokkinakis^{a)}

Department of Speech-Language-Hearing, University of Kansas, 1000 Sunnyside Avenue, Lawrence, Kansas 66045, USA

Christina Runge

Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, 9200 West Wisconsin Avenue, Milwaukee, Wisconsin 53226, USA

Qudsia Tahmina and Yi Hu

Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee, 3200 North Cramer Street, Milwaukee, Wisconsin 53211, USA

(Received 16 December 2014; revised 19 May 2015; accepted 28 May 2015; published online 6 July 2015)

The smearing effects of room reverberation can significantly impair the ability of cochlear implant (CI) listeners to understand speech. To ameliorate the effects of reverberation, current dereverberation algorithms focus on recovering the direct sound from the reverberated signal by inverse filtering the reverberation process. This contribution describes and evaluates a spectral subtraction (SS) strategy capable of suppressing late reflections. Late reflections are the most detrimental to speech intelligibility by CI listeners as reverberation increases. By tackling only the late part of reflections, it is shown that users of CI devices can benefit from the proposed strategy even in highly reverberant rooms. The proposed strategy is also compared against an ideal reverberant (binary) masking approach. Speech intelligibility results indicate that the proposed SS solution is able to suppress additive reverberant energy to a degree comparable to that achieved by an ideal binary mask. The added advantage is that the SS strategy proposed in this work can allow for a potentially real-time implementation in clinical CI processors. © 2015 Acoustical Society of America.

<http://dx.doi.org/10.1121/1.4922331>

[ELP]

Pages: 115–124

I. INTRODUCTION

Acoustic reverberation is a common phenomenon observed in nearly all enclosed spaces. It is defined as the sum total of all sound reflections arriving at a certain point inside an acoustical enclosure after the enclosure has been excited by an impulsive sound signal (Kuttruff, 2000). The reverberation time (RT_{60}) is estimated as the interval in which the reverberating sound intensity, due to decaying reflections, reaches one millionth of its initial value. In other words, it is the time it takes for the reverberation level to drop by 60 dB below the original sound energy present in the room at a given instant. For human ears or microphones that are physically placed at a relatively large distance from the target sound source, additive reverberant energy can cause significant deterioration in speech intelligibility. Many investigators have reported that for normal-hearing (NH) listeners, individuals suffering from sensorineural hearing impairment, and cochlear implant (CI) listeners, the detrimental effects of reverberation on speech perception exhibit a common pattern. In all of these populations, the intelligibility of reverberated speech progressively declines with a linear increase in reverberation time, RT_{60} (Nábělek and Pickett, 1974; Nábělek and Letowski, 1985, 1988; Nábělek

et al., 1989; Kjellberg, 2004; Neuman *et al.*, 2010; Kokkinakis *et al.*, 2011; Hazrati and Loizou, 2012; Desmond *et al.*, 2014; Hu and Kokkinakis, 2014; Mason and Kokkinakis, 2014). In this contribution, we propose a strategy to ameliorate the negative effects of reverberation on speech.

Prior studies have suggested that reverberation flattens formant transitions in vowels, results in weak-energy speech units being masked by preceding segments with strong energies, smears spectral cues, reduces temporal amplitude modulations, and increases low-frequency energy, which consequently masks high-frequency components (Bolt and MacDonald, 1949; Nábělek and Letowski, 1988; Nábělek *et al.*, 1989; Assmann and Summerfield, 2004). The study by Bolt and MacDonald (1949) was, perhaps, the first one to propose and also promote the theory of speech masking by acoustic reverberation. By assuming that a speech signal consists of a series of discrete pulses of acoustic energy separated by relatively quiet gaps, the authors computed the articulation index by measuring the intelligibility of reverberated speech by focusing mainly on overlap-masking effects. These effects were introduced by various energy components of a pulse being masked by those of previous pulses. Bolt and MacDonald (1949) assumed that the leading edge of the pulse carried the majority of the intelligibility.

^{a)}Electronic mail: kokkinak@ku.edu

The study conducted by Bolt and MacDonald (1949) suggested that self-masking effects (energy smearing within a phoneme) would have little to no effect on speech intelligibility. The hypothesis was that self-masking was mainly caused by the energy components of the leading edge of the pulse masking the remaining portions of the same pulse, which carried only minimal intelligibility in the first place. By essentially emphasizing overlap-masking effects over self-masking effects, the model yielded percent articulation values which agreed precisely with values determined experimentally at reverberation times of < 2 s. In Nábelek *et al.* (1989), using /s/-spectrum-shaped noise, the authors compared the identification of consonants following /s/ under reverberated conditions with the identification of the same consonants without the preceding /s/ phoneme under noise masking conditions. The reasoning behind this was that if there were no self-masking effects, the error patterns caused by reverberation should be comparable to those caused by the noise masker. As this was not the case, the authors ultimately concluded that both overlap-masking and self-masking had a negative impact on consonant identification.

Compared to the literature on the effects of reverberation on speech perception by NH listeners and listeners with sensorineural hearing loss, far fewer studies have examined the intelligibility of reverberated speech perceived by CI listeners. The studies by Poissant *et al.* (2006) and Whitmal and Poissant (2009) were among the first to investigate how the CI processing affects speech perception in reverberation by presenting NH adult listeners with reverberant stimuli processed into 6–24 channels using tone-excited vocoders. Poissant *et al.* (2006) showed that when only a small number of spectral bands is used to vocode speech signals inside a simulated reverberant field, the percent correct recognition scores deteriorate significantly. Performance was found to be the lowest when reverberated speech was vocoded using 6 channels, while speech recognition was only mildly affected by RT_{60} and the source-to-listener distance when the reverberant stimuli were vocoded into 12 or more channels (Whitmal and Poissant, 2009). More recently, Kokkinakis *et al.* (2011) assessed the sentence recognition performance of six adult listeners wearing CI devices using varying degrees of reverberation. This study was the first to show that speech intelligibility by individuals who relied exclusively on CIs decreased exponentially with a linear increase in reverberation time. These findings were also later corroborated by results obtained from Hazrati and Loizou (2012), who tested speech intelligibility by CI listeners in two moderate-to-severe reverberation conditions ($RT_{60} = 0.6$ s and 0.8 s). In short, both studies showed that the intelligibility of electrically processed sentences may drop by ~ 60 percentage points or more on average when reverberation time increases from $RT_{60} = 0$ s to $RT_{60} = 0.8$ s.

Speech degradation due to reverberation has led many researchers to investigate dereverberation algorithms for speech enhancement. Dereverberation often relies on inverse filtering and, hence, seeks to “invert” the reverberation process. This was one of the first and remains one of the most commonly used methods today. However, the main drawbacks of this technique include assumed prior knowledge of

the room acoustic impulse response and the sensitivity to the changes in the overall impulse response caused by even slight sound-source movements (e.g., see Kokkinakis and Loizou, 2009). In sound processing strategies used in CIs, recent efforts have focused toward developing algorithms that are not based on inverse filtering to suppress reverberation in CI devices (e.g., see Kokkinakis *et al.*, 2011; Hazrati *et al.*, 2013; Hazrati and Loizou, 2013). Kokkinakis *et al.* (2011) developed and tested an ideal reverberant masking (IRM) approach based on the estimation of the signal-to-reverberant ratio (SRR) of the individual spectral channels. The underlying motivation for using the SRR criterion was to retain the speech components arising from the early reflections, while discarding the speech components generated from the late reflections. The proposed strategy was found to yield consistent gains in intelligibility (30–55 percentage points) by CI listeners when tested with sentences from the Institute of Electrical and Electronics Engineers (IEEE) test battery (IEEE, 1969). Shortly after, Hazrati *et al.* (2013) extended the ideal binary masking approach developed in Kokkinakis *et al.* (2011) into a blind one by computing binary masks without prior knowledge of the SRR. These masks were derived by extracting a variance-based feature from the reverberant signal and comparing it against an adaptive threshold. Further, by using linear prediction analysis of the reverberant signal, Hazrati and Loizou (2013) developed an efficient reverberation suppression strategy by employing the residual-to-reverberant ratio as the channel-selection criterion. All of these methods resulted in significant intelligibility improvement of reverberated speech perceived by CI listeners.

Most dereverberation algorithms focus on recovering the direct sound from the reverberated signal. Although this is adequate for machine applications, this approach has limitations for human listeners. The reverberant sound in any enclosed space reaching the listener’s ears consists of the direct sound, early reflections, and late reflections. The direct sound travels from the sound source directly to a listener and is, generally, not influenced by the properties of the enclosure. Early reflections arrive at the listener within approximately the first 50 ms after the direct sound and are considered beneficial to the overall speech intelligibility, as they can be perceptually integrated (fused) with the direct sound (Kuttruff, 2000; Bradley *et al.*, 2003; Hu and Kokkinakis, 2014). Late reflections, on the other hand, consist of a dense succession of echoes with diminishing intensity and typically arrive at the listener with a much longer delay after the arrival of the direct sound component. These late reflections have been shown to produce mainly overlap-masking effects due to the temporal overlap of reverberant energy from a preceding phoneme onto the following phoneme (Habets, 2007; Kokkinakis *et al.*, 2011; Desmond *et al.*, 2014; Hu and Kokkinakis, 2014). In contrast, early reflections are often associated with moderate self-masking (coloration) effects due to the internal smearing of energy within each phoneme (Kokkinakis and Loizou, 2011). Given that early and late reflections will likely produce different perceptual effects on the human auditory system (Barron and Marshall, 1981; Griesinger, 1997), reverberation suppression

strategies need to be able to efficiently differentiate between these two types of reverberated (corrupted) speech.

This work is motivated by the Kokkinakis *et al.* (2011) and Hazrati *et al.* (2013) studies. The major difference between the present study and the strategies developed in Kokkinakis *et al.* (2011) and Hazrati *et al.* (2013) is that in this contribution we focus explicitly on the use of spectral subtraction (SS). SS can operate directly on the reverberant signal envelopes, unlike binary time-frequency (T-F) masking strategies, which compare the energy of the reverberant speech with that of the anechoic speech in order to determine whether to retain or discard a T-F unit. Herein, we propose a SS strategy to reduce late reverberations in speech perceived by CI listeners. Our rationale is twofold: (1) a recent study showed that it is mainly late reflections that account for the worsened speech intelligibility observed in CI listeners (Hu and Kokkinakis, 2014) and, hence, we can target only reverberant energies caused by these late reflections; and (2) SS is an efficient alternative to binary T-F strategies for reducing the masking effects caused by additive reverberant energy. By modeling the overlap-masking effects caused by late reflections as an uncorrelated random noise process, we investigate whether CI listeners could benefit from SS. The advantage is that the SS strategy proposed in this work may allow for real-time implementation in clinical CI processors.

II. STRATEGIES FOR SUPPRESSING REVERBERATION

A. IRM strategy

In the IRM strategy, both reverberant and clean signals are decomposed into T-F units. At each T-F unit, the SRR, which is equal to the ratio of the short-time energy of the clean signal to that of the reverberant signal, is computed separately for each channel. The ideal reverberant mask for that unit is subsequently obtained by comparing the SRR value against a preset threshold (T') expressed in dB, as follows (Kokkinakis *et al.*, 2011):

$$\text{IRM}(t, k) = \begin{cases} 1, & \text{SRR}(t, k) > T' \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where t corresponds to the time-frame index and k defines the frequency or channel index. We refer to this reverberation suppression strategy as the IRM strategy due to the fact that its operation assumes *a priori* knowledge of the original (uncorrupted) acoustic information.

B. SS strategy

1. Hypothesis

Our earlier findings provided strong evidence that reverberated speech caused by the late part of additive reverberation is much more detrimental to overall intelligibility than the negative impact of early reflections (Hu and Kokkinakis, 2014). In the present study, we examine an efficient spectral-subtraction based strategy for reverberation suppression, which aims to remove the late reverberation components

from the room impulse response (RIR) filter. Unlike, binary T-F masking strategies, which tend to compare the energy of the reverberant speech with that of the anechoic speech in order to determine whether to retain or discard a T-F unit (see Sec. II A), the proposed SS strategy can operate directly on the reverberant signal envelopes. The central hypothesis is that speech intelligibility improves significantly by reducing the energy due to late reverberation components through SS processing.

2. Algorithm overview

A statistical model of the RIR filter was first developed by Polack (1988). The application of this model to machine speech recognition scored modest success (Lebart and Boucher, 2001). In the current study, we use this model in the discrete-time domain to investigate whether a reverberation suppression strategy based on SS can improve the perception of reverberated speech by CI listeners. Let $s[n]$ denote the clean discrete-time speech signal, and $h[n]$ denote the RIR filter. Then, the reverberated speech signal, $x[n]$, is obtained by

$$x[n] = s[n] \otimes h[n], \quad (2)$$

where $\otimes$ indicates the discrete-time convolution operator. The causal RIR filter represented by $h[n]$ can be decomposed into three separate components: $h[0]$, which represents the direct path, $h_e[n]$, which represents the early reflection, and $h_l[n]$, which represents the late reverberation. A simplified variation of the statistical model for the RIR filter in Polack (1988) can be described as a random process with an exponentially decayed envelope signal

$$h[n] = \begin{cases} 0, & \text{for } n < 0 \\ h[0], & \text{for } n = 0 \\ h_e[n], & \text{for } 1 \leq n \leq T_e f_s - 1 \\ h_l[n], & \text{for } T_e f_s \leq n \leq N - 1, \end{cases} \quad (3)$$

$$= \begin{cases} 0, & \text{for } n < 0 \\ h[0], & \text{for } n = 0 \\ \varepsilon[n] e^{-\nu(n/f_s)}, & \text{for } 1 \leq n \leq N - 1, \end{cases} \quad (4)$$

where f_s denotes the sampling frequency, T_e refers to the duration for the early reflection part, and N indicates the length of the finite impulse response filter. $\varepsilon[n]$ is a random-variable sequence of independent and identical normal distribution with a mean of zero and a standard deviation of σ_0^2 , and the decay factor, ν , is determined by the reverberation time, RT_{60} , as

$$\nu = \frac{3 \ln(10)}{\text{RT}_{60}}. \quad (5)$$

It can be seen that the power envelope of $h[n]$ is equal to

$$E\{h^2[n]\} = \sigma_0^2 e^{-2\nu(n/f_s)}, \quad (6)$$

where $E\{\cdot\}$ denotes the ensemble averaging. By using Eqs. (2) and (4), $x[n]$ can be expressed as the sum of three parts

$$x[n] = h[0]s[n] + \underbrace{\sum_{k=1}^{T_{ef}s-1} s[n-k]h_e[k]}_{x_e[n]} + \underbrace{\sum_{k=T_{ef}s}^{N-1} s[n-k]h_l[k]}_{x_l[n]}, \quad (7)$$

where $h[0]s[n]$ is the attenuated speech signal from the direct path, $x_e[n]$ is the reverberated speech from the early reflection part of the impulse response, and $x_l[n]$ is the reverberated speech arising from the late reverberation. Alternatively, Eq. (7) can be expressed as

$$h[0]s[n] + x_e[n] = x[n] - x_l[n]. \quad (8)$$

The goal of most dereverberation algorithms is to recover the original speech signal, $s[n]$. For the task of human speech recognition, this goal usually complicates the algorithm design. This is because unlike machines, human listeners can use the early reflection part of the signal, $x_e[n]$, to facilitate recognition of reverberated speech (Hu and Kokkinakis, 2014). Based on this observation, the short-time reverberation-suppressed speech signal, $\hat{s}[j, n]$, for frame j can be estimated from the short-time reverberated speech signal, $x[j, n]$, and the short-time late-reverberation signal, $x_l[j, n]$, as

$$\begin{aligned} \hat{s}[j, n] &= h[0]s[j, n] + x_e[j, n] = x[j, n] - x_l[j, n], \quad \text{or} \\ x[j, n] &= x_l[j, n] + \hat{s}[j, n], \end{aligned} \quad (9)$$

or equivalently in the frequency domain

$$\begin{aligned} \hat{S}(j, p) &= X(j, p) - X_l(j, p), \quad \text{or} \\ X(j, p) &= X_l(j, p) + \hat{S}(j, p), \end{aligned} \quad (10)$$

where $\hat{S}(j, p)$, $X(j, p)$, and $X_l(j, p)$ denote the Fourier transform at frequency index p of $s[j, n]$, $x[j, n]$, and $x_l[j, n]$, respectively. Despite the fact that Eqs. (9) and (10) depict a straightforward approach toward estimating the clean speech signal, in practice, we have no access to the instantaneous short-time signal $x_l[j, n]$ [or equivalently $X_l(j, p)$] that is generated by a random process. Due to this, $X_l(j, p)$ will have to be estimated from the reverberated speech signal, $x[j, n]$ [or equivalently $X(j, p)$], in the statistical sense. From Eq. (9) it is evident that if $x_l[j, n]$ and $\hat{s}[j, n]$ are uncorrelated random signals, then the auto-correlation functions $r_{xx}(j, \tau)$, $r_{x_l x_l}(j, \tau)$, and $r_{\hat{s}\hat{s}}(j, \tau)$ of $x[j, n]$, $x_l[j, n]$, and $\hat{s}[j, n]$, respectively, should satisfy

$$\begin{aligned} r_{xx}(j, \tau) &= r_{x_l x_l}(j, \tau) + r_{\hat{s}\hat{s}}(j, \tau) \quad \text{or} \\ r_{\hat{s}\hat{s}}(j, \tau) &= r_{xx}(j, \tau) - r_{x_l x_l}(j, \tau), \end{aligned} \quad (11)$$

This turns out to be a justified approximation (Habets, 2007) considering that (1) the short-time speech signal, $s[j, n]$, can be assumed stationary during a time span between 20 ms and ~40 ms, and (2) the early and late components, $h_e[n]$ and $h_l[n]$, respectively, of the RIR filter are uncorrelated random signals. By observing that the reverberated speech signal, $x_l[j, n]$, from late reverberation is obtained after the convolution between $s[j, n]$ and the

exponentially decayed signal, $h_l[n]$, intuitively, we can see that the auto-correlation function, $r_{x_l x_l}(j, \tau)$, is an exponentially decayed and delayed version of $r_{xx}(j, \tau)$. Indeed, it can be shown that if $T_e \ll RT_{60}$,

$$r_{x_l x_l}(j, \tau) = e^{-2\nu T_e} r_{xx}\left(j - \frac{T_{ef}s}{R}, \tau\right), \quad (12)$$

where R is the frame rate in samples. Now, if we consider the counterpart of Eqs. (11) and (12) in terms of their power spectral densities, we can write

$$P_{\hat{s}\hat{s}}(j, p) = P_{xx}(j, p) - P_{x_l x_l}(j, p) \quad (13)$$

and

$$P_{x_l x_l}(j, p) = e^{-2\nu T_e} P_{xx}\left(j - \frac{T_{ef}s}{R}, p\right). \quad (14)$$

From the above, it quickly becomes evident that Eq. (13) is similar to the classic noise reduction method based on SS proposed by Boll (1979). Simply put, SS relies on obtaining an accurate estimate of the noise spectrum that can be subtracted from the noisy speech spectrum in order to yield a noise-attenuated speech signal. Spectral-subtraction based noise reduction strategies have been applied to CIs with much success (e.g., see Hu and Loizou, 2007). In the case of reverberant energy, the spectral variance of the estimated clean speech signal, $\hat{s}[j, n]$, can be calculated by subtracting the spectral variance of $x_l[j, n]$ from that of $x[j, n]$. Despite being a very efficient method suitable for real-time implementation, due to inaccurate power spectral density estimates, SS often suffers from an artifact called “musical noise.” A number of variations have been developed to ameliorate this problem. In the present study, we investigate one that combines the magnitude spectral over-subtraction and the spectral flooring (Berouti *et al.*, 1979), which showed the most promising results in our pilot study with one CI subject. More specifically, as shown in Fig. 1, the magnitude spectrum, $|\hat{S}(j, p)|$, can be computed by

$$|\hat{S}(j, p)| = \max[|X(j, p)| - \alpha|\hat{X}_l(j, p)|, \beta|\hat{X}_l(j, p)|], \quad (15)$$

where $\max[\cdot]$ denotes the operator to retrieve the maximum value, $|\cdot|$ indicates the absolute-value operator, α is the subtraction factor ($\alpha \geq 1$), and β is a small constant for the spectral floor. The constant factor, β , is specifically used to (1) avoid the physically impossible negative magnitude spectral values and (2) account for the case where $|X(j, p)|$ is no greater than $\alpha|\hat{X}_l(j, p)|$. Based on the block diagram of Fig. 2, $|\hat{X}_l(j, p)|$ is the estimated magnitude for $|X_l(j, p)|$, which can be written as

$$\begin{aligned} \sigma_x^2(j, p) &= \eta\sigma_x^2(j-1, p) + (1-\eta)|X(j, p)|^2, \\ |\hat{X}_l(j, p)| &= e^{-\nu T_e} \sqrt{\sigma_x^2\left(j - \frac{T_{ef}s}{R}, p\right)}, \end{aligned} \quad (16)$$

where η is a smoothing factor ($0 \leq \eta < 1$) that controls how fast $|\hat{X}_l(j, p)|$ adapts to the abrupt changes in $|X(j, p)|$. The subtraction factor, α , is computed as

FIG. 1. Overview of the proposed algorithm.

$$\alpha = \begin{cases} \alpha_+, & \text{for } \text{SNR}_{\text{post}} \geq 20 \\ \alpha_0 - \text{SNR}_{\text{post}}s, & \text{for } -5 \leq \text{SNR}_{\text{post}} \leq 20 \\ \alpha_-, & \text{for } \text{SNR}_{\text{post}} \leq -5, \end{cases} \quad (17)$$

and

$$s = \frac{\alpha_- - \alpha_+}{25}, \quad \alpha_- = 20, \quad \alpha_+ = 1, \quad \alpha_0 = \alpha_+ + 20s, \quad (18)$$

where SNR_{post} is the *a posteriori* signal-to-noise ratio defined as the ratio of $|X(j, p)|^2$ to $|\hat{X}_l(j, p)|$. This varying subtraction factor, α , across frames is used to control the trade-off between speech distortion and musical noise.

III. EXPERIMENT 1. REVERBERATION SUPPRESSION BASED ON SS

A. Methods

In this experiment, understanding of reverberant speech by CI listeners was assessed with and without the proposed SS strategy.

1. Subjects and speech stimuli

Eleven postlingually deafened adult CI listeners were recruited for testing that was conducted at the University of Kansas and the University of Wisconsin–Milwaukee.

The listeners were all native speakers of American English and were paid an hourly wage for their participation. All listeners tested had a minimum of one year of experience with their CI devices, and they used their own processors during testing. Prior to data collection, this study was approved by the Human Subjects Committee of the University of Kansas in Lawrence and the University of Wisconsin–Milwaukee’s Human Research Protection Program. All subjects gave written informed consent prior to the beginning of testing. A case history interview was conducted with each subject to determine eligibility for this experiment, as well as the experiment described in Sec. IV. Only subjects who scored > 70% on the consonant-nucleus-consonant test (Peterson and Lehiste, 1962) were included in the subject population. Complete demographic information for the individuals tested is provided in Table I.

The speech stimuli used for testing were sentences from the IEEE database. Each sentence was composed of ~7–12 words. In total there were 72 lists of 10 sentences, all produced by the same male talker. All stimuli were recorded at the sampling frequency of 25 000 Hz and were downsampled to 16 000 Hz. The root-mean-square value of all sentences was equalized to the same value corresponding to ~65 dBA. Head-related transfer functions (HRTFs) recorded by Rychtarikova *et al.* (2009) were used to simulate the two acoustically reverberant conditions ($\text{RT}_{60} = 0.3$ s and 1.0 s). To obtain measurements of the HRTFs, the authors used

FIG. 2. Estimation of reverberant magnitude.

TABLE I. Demographic details of the CI listeners who participated in this study.

Subject	Gender	Age at testing (years)	Months with CI	Etiology	Device	Experiment 1	Experiment 2
CI1	F	30	38	Hereditary	Nucleus 5	•	•
CI2	F	49	30	Noise exposure	Nucleus 5	•	•
CI3	M	22	195	Unknown	Nucleus 5	•	•
CI4	M	58	65	Hereditary	Nucleus 5	•	•
CI5	F	43	24	Unknown	Nucleus 5	•	•
CI6	M	67	20	Noise exposure	Nucleus 5	•	•
CI7	M	35	68	Meniere's	Nucleus 6	•	•
CI8	M	32	120	Meningitis	AB Harmony	•	
CI9	M	18	192	Spinal meningitis	AB Harmony	•	
CI10	M	66	72	Unknown	MED-EL	•	
CI11	M	69	120	Accident	MED-EL	•	
Mean		44.5	85.8				
Standard deviation		18.5	63.2				

a CORTEX MKII manikin artificial head placed inside a rectangular reverberant room with dimensions $5.50\text{ m} \times 4.50\text{ m} \times 3.10\text{ m}$ (length $\times$ width $\times$ height) and a total volume of 76.80 m^3 . The average reverberation time of the room (averaged in one-third-octave bands with center frequencies between 125 Hz and 4000 Hz) was equal to $RT_{60} = 1.0\text{ s}$. By increasing the number of acoustic panels and by adding extra floor carpeting to the room, as well as highly absorbent rectangular acoustic boards and panels, the average reverberation time was subsequently reduced to $RT_{60} = 0.3\text{ s}$. This latter value corresponds to a standard well-dampened room and is typical of the reverberation time often encountered in small office spaces. All recordings were facilitated using identical microphones to those used in modern behind-the-ear speech processors. To generate the corrupted stimuli, the HRTFs obtained for each reverberation condition were convolved with the speech files from the IEEE test materials using standardized linear convolution algorithms in MATLAB (MathWorks, Natick, MA).

2. Procedure

All stimuli were presented to the listeners directly through the auxiliary input jack of the CI sound processor in a double-walled sound-attenuated booth (Acoustic Systems, Inc., Austin, TX). Prior to testing, each listener participated in a short practice session to gain familiarity with the task. During the practice session, the listeners adjusted the volume to reach a comfortable level. During the test session, a 10 min break was provided every hour in order to avoid fatigue. The 11 CI listeners participated in a total of 4 experimental conditions corresponding to: (a) two different reverberant conditions ($RT_{60} = 0.3\text{ s}$ and 1.0 s) and (b) two processed conditions, one using the proposed spectral-subtraction based reverberation suppression strategy and one using reverberant stimuli fed directly into the sound processors. Two IEEE lists (20 sentences) were used per condition. None of the lists used were repeated across different conditions. To minimize any order effects, the order of the test conditions was randomized across subjects. During testing, the participants were allowed to repeat the sentence once

and they were instructed to write as many of the words as they could identify on the answer sheets. The responses of each individual were scored off-line based on the number of words correctly identified. All words were scored. The percent correct scores for each condition were calculated by dividing the number of words correctly identified by the total number of words in the particular sentence list.

B. Results

The individual speech intelligibility scores for the two different reverberant conditions ($RT_{60} = 0.3\text{ s}$ and 1.0 s) tested are shown in Figs. 3(a) and 3(b). The performance was measured in terms of percentage of the words distinguished correctly by each CI user tested. As it can be seen, significant gains in word recognition scores were obtained in both reverberant stimulus conditions using the proposed SS method. As shown in Fig. 3(a), the mean speech intelligibility scores improved from an average of 59%–75% in the $RT_{60} = 0.3\text{ s}$ condition. Additionally, for $RT_{60} = 1.0\text{ s}$, the intelligibility scores were on average 25 percentage points higher when compared to the unprocessed listening condition [see Fig. 3(b)]. Two-way analysis of variance (ANOVA) (with repeated measures) using the two different reverberant listening conditions and the two processing conditions (unprocessed and processed using SS) as within-subject factors was conducted on the rationalized arcsine unit-transformed values of the speech perception scores. A Kolmogorov-Smirnov test was run to confirm the normality of the transformed percent correct scores. Note that a critical value equal to 0.05 was used as the significance level on the statistical analyses performed. The ANOVA indicated a significant effect ($F[1,10] = 46.95$, $p < 0.001$) of the processing strategy, significant effect of reverberation time ($F[1,10] = 60.54$, $p < 0.001$) and significant interaction ($F[1,10] = 10.98$, $p = 0.008$). The observed interaction is due to the fact that the improvement in performance obtained with the proposed SS strategy, relative to the baseline condition, differed for the two reverberation times tested. *Post hoc* comparisons using Tukey's honestly significant difference tests were also run to assess significant differences in scores obtained between different conditions. For $RT_{60} = 0.3\text{ s}$,

FIG. 3. Individual percent correct scores for 11 CI listeners tested on IEEE sentences using anechoic (unprocessed) acoustic inputs recorded in reverberation (white bars) and reverberant acoustic inputs processed with the proposed SS strategy (black bars), (a) $RT_{60} = 0.3$ s and (b) $RT_{60} = 1.0$ s. Error bars indicate standard deviations.

the mean scores observed in the processed condition were higher than those obtained in the unprocessed (reverberant) condition, albeit with only a marginal statistical significance ($p = 0.047$). On the other hand, the mean scores obtained in the $RT_{60} = 1.0$ s condition, indicated that speech intelligibility improved significantly ($p = 0.003$) when processing the corrupted stimuli with the SS dereverberation strategy.

IV. EXPERIMENT 2. SS VS IRM

A. Methods

This experiment was designed to compare the improvement in speech intelligibility provided by applying the ideal reverberant mask previously developed by Kokkinakis *et al.* (2011) against the improvement observed when processing with the SS strategy introduced in Sec. II B. The IRM threshold in this case was set to -5 dB, as this has been shown to be appropriate for suppressing moderate levels of reverberation (Kokkinakis *et al.*, 2011). The ideal reverberant (binary) masking strategy has been extensively utilized as a performance upper bound for dereverberation algorithms (Kokkinakis *et al.*, 2011; Hazrati *et al.*, 2013; Hazrati and Loizou, 2013).

1. Subjects and speech stimuli

The same postlingually deafened cochlear implantees tested in experiment 1 were asked back on a different day for additional testing in order to participate in experiment 2. Only the first seven CI subjects who were using Nucleus 24 devices were available to participate in experiment 2 (see Table I). All CI listeners tested used their clinical devices programmed with the advanced combination encoder (ACE) speech coding strategy (e.g., see Vandali *et al.*, 2000). ACE is the default strategy used in Nucleus devices and operates by selecting only a subset of envelopes (typically between 8 and 12) for stimulation at each cycle. The same speech material (IEEE, 1969) was used as in experiment 1. To avoid potential learning effects, none of the sentence lists previously used were reused.

2. Procedure

In this experiment the subjects participated in a total of four testing conditions corresponding to (a) reverberant stimuli ($RT_{60} = 0.3$ s and 1.0 s) processed with the IRM strategy (e.g., see Kokkinakis *et al.*, 2011) described in Sec. II A, and (b) reverberant stimuli ($RT_{60} = 0.3$ s and 1.0 s) processed with the proposed spectral-subtraction based reverberation suppression strategy (see Sec. II B). As before, each participant completed all conditions in a single test session. Two IEEE lists (20 sentences) were used for each condition. A total of 80 sentences were used in this experiment. Each sentence was presented once. The order of the test conditions was randomized across subjects. All the stimuli were presented directly through the auxiliary input jack of the CI sound processors.

B. Results

Figure 4 plots the individual subjects' scores obtained using the two aforementioned dereverberation strategies. For $RT_{60} = 0.3$ s, the average speech intelligibility scores for all listeners tested was approximately equal to 77% for the IRM condition and $\sim 72\%$ when the corrupted stimuli were processed using SS. The individual subject scores revealed that almost half of all the individuals tested did better with the SS technique in the 0.3 s condition. For $RT_{60} = 1.0$ s, the mean scores across all subjects were equal to 70% and 61% for the IRM and SS conditions, respectively. In this condition, the scores indicated that the majority of subjects tested performed better when processing the corrupted stimuli using the IRM strategy. Two-way ANOVA (with repeated measures) using the two different listening conditions and the two processing conditions (IRM- and SS-processed) as within-subject factors was conducted on the rationalized arcsine unit-transformed values of the speech perception scores. The ANOVA indicated a non-significant effect ($F[1,10] = 4.53$, $p = 0.078$) of the processing strategy, a significant effect of reverberation time ($F[1,10] = 45.84$, $p = 0.001$) and a non-significant interaction ($F[1,10] = 2.14$, $p = 0.19$) between the condition and the strategy.

FIG. 4. Individual percent correct scores obtained from seven Nucleus 24 CI listeners tested on IEEE sentences using reverberant acoustic inputs processed with (a) the IRM strategy (white bars) and (b) the proposed SS strategy (black bars), (a) $RT_{60} = 0.3$ s and (b) $RT_{60} = 1.0$ s. Error bars indicate standard deviations.

V. DISCUSSION

The average speech intelligibility scores for all CI listeners tested were equal to 59% and 35% for the two reverberant conditions ($RT_{60} = 0.3, 1.0$ s). The findings of this study are consistent with the results obtained from an earlier study by Kokkinakis *et al.* (2011), which concluded that there is a very strong, and negative, relationship between speech perception and the amount of additive acoustical reverberation. In line with the Kokkinakis *et al.* (2011) study, speech intelligibility by CI users in reverberation degraded exponentially with a linear increase in reverberation time. A number of factors contributed to the low performance observed. These include the low spectral resolution due to standard CI sound processing, self-masking effects (causing at formant transitions) that are often associated with early reflections, and the detrimental effects of temporal envelope smearing (overlap-masking). Of these three factors, we believe that the negative effects of temporal envelope smearing, introduced predominantly by overlap-masking, contributed the most. Overlap-masking results in temporal smearing, which often causes one phoneme to be masked by the reverberant energy present in a preceding phoneme. This temporal envelope smearing caused by reverberation (particularly by late reflections) makes the detection of acoustic landmarks extremely difficult, and that disrupts the syllable structure, which is known to be important for determining word boundaries (Stevens, 2002).

Figure 5 illustrates example stimulus output patterns (electrodograms) of an IEEE sentence processed with the ACE speech coding strategy, which is predominantly used in conjunction with the Nucleus 24 device. In all panels shown, the vertical axes represent the electrode position corresponding to a specific frequency, while the horizontal axes show time progression. For this example, speech was corrupted by additive reverberation in a room with $RT_{60} = 1.0$ s. Temporal envelope smearing is evident in Fig. 5(b). As shown in Fig. 5(b), temporal smearing blurs the vowel and consonant boundaries which are normally present in the anechoic stimuli plotted in Fig. 5(a). The prolonged formant transitions caused by self-masking are also evident in Fig. 5(b) (electrodes 10–15) when focusing on vowel targets (e.g., see vowel /ɔ/ at 0.30–0.45 s and vowel /e/ at 1.10–1.25 s). Such self-masking (or coloration) effects, which arise due to the internal smearing of energy within each phoneme, are believed to be either less or equally damaging to overlap-masking effects (Desmond *et al.*, 2014; Hu and Kokkinakis, 2014).

Figure 5(c) depicts the output patterns of stimuli processed using the ideal reverberant mask previously described in Kokkinakis *et al.* (2011). As it is evident from Fig. 5(c), the IRM strategy can suppress additive reverberant energy to a substantial extent. The IRM strategy relies on a local SRR criterion and can correctly discard the reverberant channels corrupted by late reflections, while selecting only those T-F channels containing primarily the signal consisting of the direct component and some early reflections. As shown in Fig. 5(b), the ACE strategy will select mainly the low- and mid-frequency channels (electrodes 10–18) instead of the high-frequency channels when coding consonant segments. This is due to the principle underlying the operation of ACE, in that speech can be understood well even if only the peaks in the short-term spectrum are transmitted. For reverberant stimuli, intense low- and mid-frequency energy present in the stimulus could capture all the available electrode-driving time slots. Thus, none will remain available to respond to the weaker high-frequency energy contained in the unvoiced part of the speech input. In contrast, the IRM strategy can correctly select the high-frequency channels in the most basal electrodes (electrodes 1–9) corresponding to unvoiced segments (e.g., consonants). Therefore, when pre-processing the corrupted (reverberant) speech stimuli with the IRM strategy, the correct bands with the highest spectra will be selected to drive the electrodes, instead of the mid-frequency channels that contain mainly unwanted additive reverberant energy.

As illustrated in Fig. 5(d), low- and mid-frequency energy due to additive acoustic reverberation can also be attenuated considerably (especially in the apical electrodes) after processing with the proposed SS strategy. In addition, the vowel and consonant boundaries, as well as silence gaps previously filled with reverberant energy are recovered resulting in improved useful acoustic cues and an increase in the observed speech intelligibility. In fact, when comparing the electrodogram of the IRM-processed reverberant speech shown in Fig. 5(c) with that of the spectral subtraction-processed reverberant signal in Fig. 5(d), it is evident that

FIG. 5. Stimulus output patterns (electrograms) of the IEEE sentence “The water in this well is a source of good health” uttered by a male speaker. (a) Electrogram of unmodified (uncorrupted) sentence processed by the ACE strategy, (b) electrogram of the same sentence when corrupted by reverberation equal to $RT_{60} = 1.0$ s and processed by the ACE strategy, (c) electrogram of the reverberant sentence when processed by the IRM speech coding strategy, and (d) electrogram of the reverberant sentence when processed proposed SS speech coding strategy. In each electrogram, time is shown along the abscissa and the electrode number is shown along the ordinate.

both strategies were able to eliminate temporal envelope smearing effects caused by overlap-masking to almost an equal extent. This is also reflected in the non-significant gap observed in the intelligibility scores between the two separate processing conditions. The advantage of the proposed strategy is that it can operate with little to almost no prior information on the signal or the room characteristics, while the IRM strategy requires estimating the SRR criterion from the uncorrupted signal envelopes. This makes the proposed strategy amenable to real-time implementation.

VI. CONCLUSIONS

In this contribution, we have developed a SS dereverberation strategy capable of suppressing reverberant energy components due to late reflections. This strategy was evaluated using CI listeners and two different reverberation time constants. Speech intelligibility scores observed in the processed conditions suggest that the proposed solution can be used to successfully ameliorate the negative impact of late reflections on the perceived speech. Given the potential benefits, in terms of intelligibility, such a strategy could be potentially implemented in clinical CI processors for reverberation suppression in complex listening settings.

ACKNOWLEDGMENTS

This research was supported in part by a New Faculty General Research Fund grant from the University of Kansas (K.K.) and a Research Growth Initiative grant from the University of Wisconsin–Milwaukee (Y.H.). The authors would like to thank the CI patients for their time and dedication. The authors would also like to acknowledge the help of Cochlear Limited (New South Wales, Australia).

- Assmann, P. F., and Summerfield, Q. (2004). “The perception of speech under adverse acoustic conditions,” in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York), pp. 231–308.
- Barron, M., and Marshall, A. H. (1981). “Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure,” *J. Sound. Vib.* **77**, 211–232.
- Berouti, M., Schwartz, R., and Makhoul, J. (1979). “Enhancement of speech corrupted by acoustic noise,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 208–211.
- Boll, S. F. (1979). “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.* **27**, 113–120.
- Bolt, R. H., and MacDonald, A. D. (1949). “Theory of speech masking by reverberation,” *J. Acoust. Soc. Am.* **21**, 577–580.
- Bradley, J. S., Sato, H., and Picard, M. (2003). “On the importance of early reflections for speech in rooms,” *J. Acoust. Soc. Am.* **113**, 3233–3244.

- Desmond, J. M., Collins, L. M., and Throckmorton, C. S. (2014). "The effects of reverberant self- and overlap-masking on speech recognition in cochlear implant listeners," *J. Acoust. Soc. Am.* **135**, EL304–EL310.
- Griesinger, D. (1997). "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces," *Acustica* **83**, 721–731.
- Habets, E. A. P. (2007). "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische University Eindhoven, Eindhoven, The Netherlands.
- Hazrati, O., Lee, J., and Loizou, P. C. (2013). "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Am.* **133**, 1607–1614.
- Hazrati, O., and Loizou, P. C. (2012). "The combined effects of reverberation and noise on speech intelligibility by cochlear implant users," *Int. J. Audiol.* **51**, 437–443.
- Hazrati, O., and Loizou, P. C. (2013). "Reverberation suppression in cochlear implants using a blind channel-selection strategy," *J. Acoust. Soc. Am.* **133**, 4188–4196.
- Hu, Y., and Kokkinakis, K. (2014). "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *J. Acoust. Soc. Am.* **135**, EL22–EL28.
- Hu, Y., and Loizou, P. (2007). "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.* **122**, 1777–1786.
- IEEE (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 225–246.
- Kjellberg, A. (2004). "Effects of reverberation time on the cognitive load in speech communication: Theoretical considerations," *Noise Health* **7**, 11–21.
- Kokkinakis, K., Hazrati, O., and Loizou, P. C. (2011). "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.* **129**, 3221–3232.
- Kokkinakis, K., and Loizou, P. C. (2009). "Selective-tap blind dereverberation for two-microphone enhancement of reverberant speech," *IEEE Signal Process. Lett.* **16**, 961–964.
- Kokkinakis, K., and Loizou, P. C. (2011). "The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners," *J. Acoust. Soc. Am.* **130**, 1099–1102.
- Kuttruff, H. (2000). *Room Acoustics* (Taylor and Francis, New York).
- Lebart, K., and Boucher, J. M. (2001). "A new method based on spectral subtraction for speech dereverberation," *Acta Acust.* **87**, 359–366.
- Mason, M., and Kokkinakis, K. (2014). "Perception of consonants in reverberation and noise by adults fitted with bimodal devices," *J. Speech Lang. Hear. Res.* **57**, 1512–1520.
- Nábělek, A. K., and Letowski, T. R. (1985). "Vowel confusions of hearing impaired listeners under reverberant and non-reverberant conditions," *J. Speech Hear. Disord.* **50**, 126–131.
- Nábělek, A. K., and Letowski, T. R. (1988). "Similarities of vowels in non-reverberant and reverberant fields," *J. Acoust. Soc. Am.* **83**, 1891–1899.
- Nábělek, A. K., Letowski, T. R., and Tucker, F. M. (1989). "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.* **86**, 1259–1265.
- Nábělek, A. K., and Picket, J. M. (1974). "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.* **17**, 724–739.
- Neuman, A. C., Wroblewski, M., Hajicek, J., and Rubinstein, A. (2010). "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear Hear.* **31**, 336–344.
- Peterson, F. E., and Lehiste, I. (1962). "Revised CNC lists for auditory tests," *J. Speech. Hear. Disord.* **27**, 62–70.
- Poissant, S. F., Whitmal, N. A. III, and Freyman, R. L. (2006). "Effects of reverberation and masking on speech intelligibility in cochlear implant simulations," *J. Acoust. Soc. Am.* **119**, 1606–1615.
- Polack, J. D. (1988). "La transmission de l'énergie sonore dans les salles" ("The transmission of sound energy in auditoria"), Ph.D. dissertation, Université du Maine, La Mans, France.
- Rychtarikova, M., Van den Bogaert, T., Vermeir, G., and Wouters, J. (2009). "Binaural sound source localization in real and virtual rooms," *J. Aud. Eng. Soc.* **57**, 205–220.
- Stevens, K. N. (2002). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *J. Acoust. Soc. Am.* **111**, 1872–1891.
- Vandali, A. E., Whitford, L. A., Plant, K. L., and Clark, G. M. (2000). "Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 cochlear implant system," *Ear Hear.* **21**, 608–624.
- Whitmal, N. A., and Poissant, S. F. (2009). "Effects of source-to-listener distance and masking on perception of cochlear implant processed speech in reverberant rooms," *J. Acoust. Soc. Am.* **126**, 2556–2569.