The potential of onset enhancement for increased speech intelligibility in auditory prostheses

Raphael Koning[a] and Jan Wouters

Experimental Otorhinolaryngology, Department Neurosciences, KU Leuven, O & N 2, Herestraat 49 bus 721,
B-3000 Leuven, Belgium

(Received 4 May 2012; revised 10 August 2012; accepted 13 August 2012)

Recent studies have shown that transient parts of a speech signal contribute most to speech intelligibility in normal-hearing listeners. In this study, the influence of enhancing the onsets of the envelope of the speech signal on speech intelligibility in noisy conditions using an eight channel cochlear implant vocoder simulation was investigated. The enhanced envelope (EE) strategy emphasizes the onsets of the speech envelope by deriving an additional peak signal at the onsets in each frequency band. A sentence recognition task in stationary speech shaped noise showed a significant speech reception threshold (SRT) improvement of 2.5 dB for the EE in comparison to the reference continuous interleaved sampling strategy and of 1.7 dB when an ideal Wiener filter was used for the onset extraction on the noisy signal. In a competitive talker condition, a significant SRT improvement of 2.6 dB was measured. A benefit was obtained in all experiments with the peak signal derived from the clean speech. Although the EE strategy is not effective in many real-life situations, the results suggest that there is potential for speech intelligibility improvement when an enhancement of the onsets of the speech envelope is included in the signal processing of auditory prostheses.

© *2012 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4748965]

PACS number(s): 43.66.Ts, 43.71.Rt, 43.60.Hj [EB] Pages: 2569–2581

I. INTRODUCTION

The objectives of speech enhancement algorithms are to improve aspects of intelligibility or quality of noisy speech signals. One approach for speech enhancement is to reduce as much noise as possible from the noisy speech signal by applying noise reduction algorithms based on spectral subtraction (Boll, 1979), statistical modeling (Hendriks and Martin, 2007), and Wiener-filtering (Chen *et al.*, 2006) while introducing a minimum of speech distortion. A comparison between eight noise reduction algorithms (spectral subtractive, subspace, statistical-model based and Wiener-type algorithms) in terms of speech quality (Hu and Loizou, 2007a) and speech intelligibility (Hu and Loizou, 2007b) showed that there was no correlation between speech intelligibility scores and speech quality. Speech and noise distortions introduced by the processing of the noisy mixture affect the speech quality. All tested algorithms were not able to significantly improve speech intelligibility across four different noise conditions (babble, car, street, and train noise) and just six maintained speech intelligibility. It was reported that across all conditions the Wiener-type algorithm performed best. Another study by Luts *et al.* (2010) evaluated the performance of noise reduction algorithm in babble noise and obtained no speech reception threshold (SRT) improvement for all single channel noise reduction algorithms and a significant SRT improvement of 6 dB for a spatially preprocessed speech-distortion-weighted multi-channel Wiener filtering. In both studies, the algorithms that performed best in terms of speech quality were not the best in terms of speech intelligibility. A literature overview on noise reduction techniques may be found in Vary and Martin (2006) or Benesty *et al.* (2005).

Another approach to improve the intelligibility of a speech signal in adverse listening conditions is based on an intentional distortion of the signal. Motivated by the fact that the concept of the modulation transfer function in room acoustics introduced by Schroeder (1981) correlates highly with speech intelligibility (Houtgast and Steeneken, 1985) the role of compression and expansion of the temporal envelope on speech intelligibility was investigated. The relation between speech intelligibility and the modulation transfer function in adverse listening conditions is expressed by the speech transmission index. The modulation depth at all frequencies of target speech signal decreases in adverse listening conditions. Langhans and Strube (1982) studied the effect of increasing the modulation depth of corrupted speech by nonlinear multiband envelope filtering. They found that speech intelligibility increases in noisy environments when compression was performed on low modulation frequencies and expansion was done at higher modulation frequencies above 2–16 Hz.

While amplitude compression was shown to have detrimental effects on speech intelligibility for normal-hearing (NH) and hearing-impaired (HI) listeners (Plomp, 1988; Fu and Shannon, 1999; van Buuren *et al.*, 1999), moderate envelope expansion was found to be beneficial especially in adverse listening conditions (Clarkson and Bahgat, 1991; Fu and Shannon, 1999). Fu and Shannon (1999) reported a small decrease in speech intelligibility in quiet for the expansion of the envelope by a power law function. However, in noisy conditions performance increased. The envelope expansion led to perceivable speech distortion (van Buuren *et al.*, 1999). Therefore, there is an upper limit of envelope expansion that can be applied.

[a] Author to whom corresponding should be addressed. Electronic mail: raphael.koning@med.kuleuven.be

Manipulating the whole envelope by compression or expansion did not lead to the desired increases in speech intelligibility. Therefore, the development of envelope enhancement algorithms continued and the focus shifted to specific parts of the speech signal that were emphasized. There is a broad discussion about which parts of the speech signal contribute most to speech intelligibility and therefore on which parts of the speech envelope the enhancement strategies should focus.

A typical classification of a speech signal into its components differentiates between vowels and consonants. In a number of studies (Kewley-Port *et al.*, 2007; Lee and Kewley-Port, 2009; Fogerty and Kewley-Port, 2009), the contribution of vowels and consonants to speech intelligibility was investigated. They reported that the speech intelligibility was two times higher for sentences where the vowels were unprocessed and the consonants were replaced with stationary speech shaped noise (SSN) over sentences with replaced vowels and unprocessed consonants. This suggested that in NH listeners, the vowels contribute more to speech intelligibility than consonants. A larger benefit for the vowel-only sentences was even obtained in a study with elderly hearing-impaired listeners (Kewley-Port *et al.*, 2007). In contrast to the findings in these studies, Owren and Cardillo (2006) found that the consonants contribute more to word meaning identification than vowels when replacing the other word part not by SSN but by gaps of silence. They chose the silence replacement because the SSN replacement can lead to a phonetic restoration of the consonant structure which would lead to more top-down processes, especially with contextually rich sentences. They did not study the effect on sentence level.

The classification into vowels and consonants without taking transitions between stationary and non-stationary parts into account seem to be too broad and general. Especially, because it was shown that all sensorineural systems (i.e., vision, taste, touch, smell, and audition) are by nature very sensitive to changes of the input signal. Kluender *et al.* (2003) investigated the importance of changing cues in coarticulated speech. They found that the auditory system has its highest sensitivity to changes in the spectral and temporal characteristics of the speech signal. Lewicki (2010) reported that the transient parts of the speech signal contribute most to speech intelligibility. This is underlined by a study of Stilp and Kluender (2010), where they developed an instrumental measure called cochlea-scaled entropy that measures the predictability of consecutive time frames. They showed that reducing speech parts with high cochlea-scaled entropy causes a decrease in speech intelligibility. High entropy parts occur at transients, onsets, and offsets. They are characterized by rapid changes in the spectral and temporal characteristics. Chen and Loizou (2012) investigated how well speech intelligibility can be predicted if sentences are segmented based on different measures: cochlea-scaled entropy, normalized root mean square amplitude, and a sonorant/obstruent segmentation. They used the speech transmission index based normalized covariance measure for the prediction which is a good measure for speech intelligibility in HI listeners (Goldsworthy and Greenberg, 2004). The prediction with the normalized covariance measure was best when mid-level root mean squares segments were evaluated. The latter outperformed the prediction with the high entropy parts of the cochlea-scaled entropy segmentation. They conclude that this is based on the fact that the mid-level root mean squares segments consist of consonant-vowel (CV) and vowel-consonant (VC) transitions. High cochlea-scaled entropy occurs also in formant transitions within vowels but do not contribute much to speech intelligibility. They also conclude that a problem of the cochlea-scaled entropy measure is that the time constant of 80 ms is too large to react on rapid changes of the input envelope and that stationary parts of the vowel leak into high entropy parts.

Some studies with NH listeners investigated the importance of transient cues on speech intelligibility. The influence of the transitions between consecutive phonemes on speech intelligibility was demonstrated in a study of Strange *et al.* (1983). They showed that consonant-vowel-consonant (CVC) syllables of which the stationary part of the vowel was removed, but still contained the CV and the VC transitions, was as intelligible as the complete CVC syllable.

The contribution of consonant landmarks in acoustic-electric hearing was investigated in a study by Chen and Loizou (2010). They obtained a 30% improvement in speech recognition, when the listener had access to the clean obstruent consonants up to 600 Hz. The rest of the signal was corrupted at −5 and 0 dB signal-to-noise ratio (SNR) in two-talker and steady-state noise. The part of the signal in the frequency range above 600 Hz was vocoded. In a second experiment, the obstruent consonants were left corrupted but they were attenuated. Therefore, the listeners had access to the boundaries of the phonemes that occur at the onsets and offsets of the signal. They reported an increase of 14% in the speech intelligibility score in the second experiment.

Most of the enhancement strategies following the latter approach for NH listeners are therefore focused on the onsets and transients in speech. It is even more crucial to focus on these speech parts because the dynamic changes can easily be affected in adverse listening conditions. A number of studies (Hazan and Simpson, 1998; Lorenzi *et al.*, 1999; Apoux *et al.*, 2004; Skowronski and Harris, 2006; Yoo *et al.*, 2007; Rasetshwane *et al.*, 2009) investigated the effect of increasing the CV ratio or amplifying the transient parts of the target speech signal with NH listeners. Amplifying the transient parts lead to an improvement in speech intelligibility in various noisy conditions. Kennedy *et al.* (1998) reported that increasing the CV intensity ratio had a significant effect on the consonant recognition scores in CV words. The maximum performance was achieved when the amplification factor was adjusted for each subject individually.

All the applied algorithms depend on a manual extraction of the peak or a complex algorithm that at present would not allow a real-time implementation. Furthermore, most of them were developed under the assumption that the clean speech signal is available and the signal is also enhanced before it is mixed with the noise background.

Only few algorithms have been proposed that focus on the transient parts of the signal for cochlear implants (CIs).

Vandali (2001) developed the transient emphasis spectral maxima (TESM) strategy that amplifies transient parts of the signal with a gain factor derived from a comparison of the averaged energy in three consecutive time windows in each processing channel. In several studies with CI listeners (Vandali, 2001; Bhattacharya *et al.*, 2011), it was shown that small but significant improvements in speech intelligibility were obtained with the TESM strategy in quiet for consonant recognition and in multitalker babble noise at 5 dB SNR (Vandali, 2001) or in combination with a spectral expansion stage (Bhattacharya *et al.*, 2011). In contrast, Holden *et al.* (2005) found no increase in speech intelligibility with this strategy.

Another strategy that was particularly focused on the onset of the speech envelope is the enhanced envelope continuous interleaved sampling (EECIS) strategy (Geurts *et al.*, 1999). The EECIS strategy was developed based on the idea of mimicking the short-term temporal adaptation characteristics of the auditory nerve synapse that is bypassed by CI electrical stimulation. The rapid adaptation effect of the auditory nerve synapse results in a higher discharge probability of neuro-transmitter at onsets of the signal (Delgutte and Kiang, 1984). In each frequency band, the algorithm detects and amplifies onsets in each frequency band. While the TESM strategy derives a gain factor that is not constant during the amplification of a transient and reaches its maximum value at the peak of the transient, the EECIS strategy provides an almost constant gain factor at the detected onset of the speech envelope. The EECIS strategy is focused on the onsets of the speech envelope in each frequency band whereas the TESM strategy also enhances transients that occur within one frequency band without a speech pause in between. It was shown in CI listeners that the place of articulation consonant feature was better transmitted in quiet and a small but significant improvement in speech intelligibility was obtained in a vowel-consonant-vowel identification task. The algorithm was neither tested in noise nor on the sentence level.

Speech enhancement algorithms are mostly developed for NH and hearing-impaired listeners. The feasibility study of the algorithms for application in CI recipients is often carried out in a first stage using vocoder simulations as a model of CI processing and evaluated with NH listeners.

The focus of this study is to investigate if enhanced onset cues provided by the proposed enhanced envelope (EE) strategy can improve speech intelligibility with noise vocoded speech in different interfering background sounds. The EE strategy is based on the EECIS strategy and amplifies the onsets of the speech envelope in each frequency band leading to a sharp onset without affecting other parts of the speech signal. In contrast, none of the mentioned studies in NH listeners were focused on the effect of amplified onsets but more on enhanced transient sounds. Also the TESM strategy for CIs provides the highest gain to the maximum amplitude of the transient while the EE strategy amplifies the onsets in the speech envelope. The EE strategy fulfills the constraint that the complexity of the algorithm is low and a real-time implementation would be possible.

The enhancement of onset cues in noise vocoded speech was investigated with three different sentence recognition tasks. The EE strategy was investigated in comparison to the reference continuous interleaved sampling (CIS) strategy in stationary SSN and in a competing talker condition. The noisy maskers were chosen to investigate the effect on a stationary and a highly non-stationary masker. The competing talker situation was included because CI users suffer from a decrease in speech intelligibility performance in fluctuating noisy maskers (Nelson *et al.*, 2003) and it is one of the most challenging situations for single channel speech enhancement strategies (Bronkhorst, 2000). The influence of the envelope enhancement under ideal peak extraction conditions in SSN is investigated in the first experiment. The performance of the approach in a real-time application is assessed in the second experiment with the introduction of a front-end Wiener filter processing step to extract the onsets from the noisy input signal. The influence of enhanced onsets on speech intelligibility in the two talker condition is investigated in the third experiment.

We hypothesized that the access to onset cues is crucial in adverse listening conditions. Onset cues play an important role in source segregation in auditory scene analysis. Common onsets and offsets across frequency can be used to separate different sound sources (Bregman, 1990) because it is very unlikely that the target signal and the interfering background are modulated coherently. Due to these characteristics, computational auditory scene analysis approaches with grouping based on onsets and offsets were developed (Hu and Wang, 2007). Emphasizing the common onsets could lead to a better segregation of the target signal and the interfering background. Shamma *et al.* (2011) demonstrated that stream formation is primarily based on temporally coherent features of the target signal. Moreover, pointing the attention to one feature can help to segregate better the target from the interfering background. Onset cues are likely to be affected in noisy listening conditions. Therefore, it is clear that the enhancement of these cues may help to segregate the target sound from the interferer. Especially for noise vocoded speech, less cues are available for segregation because temporal fine structure information is missing (Shannon *et al.*, 1995).

II. METHODS

The impact of enhancement of onsets in the speech envelope on speech perception in noisy environments was investigated in NH listeners using vocoder simulations as a model for CI speech processing. Four different listening tasks were used to evaluate the signal processing: three sentence recognition tasks, two in stationary SSN and one with an interfering talker, and a loudness rating (LR) task. The LR task was conducted to evaluate the influence of the signal processing on the loudness perception to control for possible loudness effects.

A. Subjects

Three groups of ten NH listeners participated in the different listening tasks. The first group participated in the first sentence recognition task in stationary SSN and in the LR task. Another group participated in the second sentence recognition task in stationary SSN with the envelope enhancement applied to the noisy mixture. The third group of ten participants conducted a sentence recognition task in a competing talker condition.

All 30 subjects were native speakers of Dutch/Flemish. They all had hearing thresholds below 20 dB hearing level on the octave frequencies between 125 Hz to 8 kHz. The subjects participated voluntarily in the experiments and signed an informed consent form. None of the subject were acquainted with the used speech material before conducting the listening tasks.

B. Signal processing

The EE strategy is developed based on the EECIS strategy that was originally developed for CI (Geurts *et al.*, 1999). In this study, the EE strategy was evaluated in comparison to the reference CIS strategy using vocoder simulations as a model for simulation of CI signal processing.

Both speech processing strategies consist of the following stages: bandpass filtering, envelope extraction and vocoding. The enhancement of the onsets of the speech envelope is done in the envelope extraction stage. The standard envelope $E_{\mathrm{CIS}}(\lambda, k)$, where λ represents the discrete frame index and k the channel index, of the CIS strategy has no additional envelope enhancement step in the envelope extraction step.

The input signal is sampled with a rate of 16 kHz and split into frames with a length of 128 samples and a frame advance of 16 samples. The bandpass filtering is performed by a weighted sum of the power of frequency bins that are obtained by calculating the fast Fourier transform of the sample window. The analysis window is a cosine window. After calculating the power in each frequency bin, the frequency bins are summed with a weighting factor to map the frequency bins to eight channels. The cutoff frequencies for the eight channels are 187.5, 437.5, 687.5, 1062.5, 1562.5, 2312.5, 3437.5, 5187.5, and 7937.5 Hz, which corresponds to bandwidths of 250, 250, 375, 500, 750, 1125, 1750, and 2750 Hz. All signal parts under 187.5 Hz are not considered in the processing. All settings are the default settings that are used in the CI devices of Cochlear, Ltd. for the CIS strategy. The number of channels was chosen to be representative of the CIS strategy and of the advanced combination encoder (ACE) strategy in CI. The ACE strategy performs a n-of-m maxima selection, where n is the number of selected channels and m is the total amount of available channels. It was shown before by Friesen *et al.* (2001) that asymptotic speech intelligibility performance for most CI users is reached when the number of effective channels is 8. Therefore, often eight channels are also selected in the maxima selection for CI users.

The EE strategy enhances the onsets of the speech envelope in each channel. The envelope enhanced signal is the sum of the standard envelope of the bandpass filtered signal and a peak signal that is extracted at each sudden increase in energy in the envelope by a comparison between the standard envelope and a so called "slow" envelope.

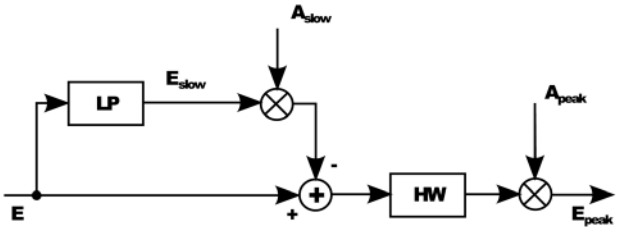

FIG. 1. Derivation of the peak signal E_{peak} at onsets of an envelope E in one frequency band by a comparison between an amplified low-pass filtered envelope E_{slow} and the input envelope. The block LP represents a 4th order Butterworth low-pass filter with a cutoff frequency of 20 Hz. The amplification factor A_{slow} is used to amplify E_{slow} to an extent that the envelope has higher amplitudes than E for the quasi stationary part of the signal. The half-wave rectification of the peak signal after the comparison of the two envelopes is represented by the block HW. The resulting envelope is amplified by the factor A_{peak} to obtain the final peak signal E_{peak}.

The algorithm to extract the additional peak signal is shown in Fig. 1. Due to the fact that the input signal to the peak extraction stage differs in the experiments, a general description of the derivation of the additional peak signal is given and the input envelope is called $E(\lambda, k)$. The slow envelope $E_{\mathrm{slow}}(\lambda, k)$ is obtained by filtering $E(\lambda, k)$ using a fourth order Butterworth low-pass filter with a cutoff frequency of 20 Hz represented by the block named LP in Fig. 1. Due to the low cutoff frequency, no F0 modulation is present in the slow envelope. Additionally, it has a bigger time delay in reacting to sudden increases of $E(\lambda, k)$. The slow envelope is amplified by a factor of $A_{\mathrm{slow}} = 8$ to ensure that the level of $E_{\mathrm{slow}}(\lambda, k)$ at quasi-stationary parts of the signal is higher than of $E(\lambda, k)$. At sudden increases in energy, $E(\lambda, k)$ lies above the slow envelope $E_{\mathrm{slow}}(\lambda, k)$ and this part is extracted as the peak signal by subtracting the slow envelope $E_{\mathrm{slow}}(\lambda, k)$ from $E(\lambda, k)$ followed by a half-wave rectification (HW in Fig. 1). The half-wave rectification is required, because the subtraction results in negative values of the peak signal at speech parts at temporarily stationary levels. An amplification factor of $A_{\mathrm{slow}} = 8$ is used to ensure that just the onsets are detected. The value was chosen that there is no leakage of the stationary part in the extracted peak envelope for the Leuven Intelligibility Sentence Test (LIST) sentences. The factor was also tested for the VU (Versfeld *et al.*, 2000) and the BKB (Bench *et al.*, 1979) sentences and no leakage occurred with this value of $A_{\mathrm{slow}} = 8$. Therefore, the peak signal has just values different from zero at the onsets of the envelope in the respective channel. The peak signal is amplified by a factor of $A_{\mathrm{peak}} = 6$ to obtain the final peak signal $E_{\mathrm{peak}}(\lambda, k)$. The derivation of the peak signal in the k-th channel can be finally written as

$$E_{\mathrm{peak}}(\lambda, k) = A_{\mathrm{peak}} \max\Big(E(\lambda, k) - A_{\mathrm{slow}} E_{\mathrm{slow}}(\lambda, k), 0\Big). \tag{1}$$

The enhanced envelope $E_{\mathrm{EE}}(\lambda, k)$ of the EE strategy is obtained by adding the peak signal $E_{\mathrm{peak}}(\lambda, k)$ to the standard CIS envelope $E_{\mathrm{CIS}}(\lambda, k)$

$$E_{\mathrm{EE}}(\lambda, k) = E_{\mathrm{CIS}}(\lambda, k) + E_{\mathrm{peak}}(\lambda, k). \tag{2}$$

The effect of the envelope enhancement is only present at sudden increases in the power of the envelope in each channel.

The value of the amplification of the peak signal was chosen as maximal without the occurrence of clipping for all sentences of the LIST material (van Wieringen and Wouters, 2008).

In the original EECIS version (Geurts *et al.*, 1999), the peak signal was extracted after an envelope compression step. Extracting the peak after the compression had the disadvantage that the algorithm was not sensitive to the low energy parts of the speech signal and therefore they were lost in the compression stage and not considered in the peak extraction step. In the new version of the algorithm, the peak signal is extracted from the uncompressed envelope. This has the advantage that the weak parts at the onsets of the signal are detected and therefore amplified more appropriately.

In the synthesis stage, the envelope of the CIS strategy and the modified envelopes of the EE strategy were used to modulate a broadband noise carrier (Shannon *et al.*, 1995). The modulated noise was filtered by the same filter bank that was used in the analysis stage. All noise vocoded channels are then added to obtain the final stimulus of the speech signal.

Due to the fact that the input signal to the peak signal extraction stage was different in the three different speech recognition tasks, the signal processing for the different enhanced conditions is briefly introduced here.

In experiment I, the potential impact of the enhancement algorithm was evaluated when speech and noise components were independently processed. Therefore, the input envelope of the peak extraction stage was the clean speech signal that was enhanced with the extracted peak signal and afterwards mixed at the desired SNR with the separately vocoded SSN. For the reference CIS condition, the mixing was also done with the separately vocoded speech and noise components.

In Fig. 2, the clean speech signal (above) and the vocoded output of the EE strategy are shown for the Dutch sentence "Morgen gaan we naar de stad (Tomorrow we are going to the city)." Note that just the onsets are amplified by the EE algorithm and the stationary part of the signal is not affected in comparison to the CIS strategy. This enhanced signal is then mixed afterwards with the vocoded noise signal at the desired SNR in experiment I.

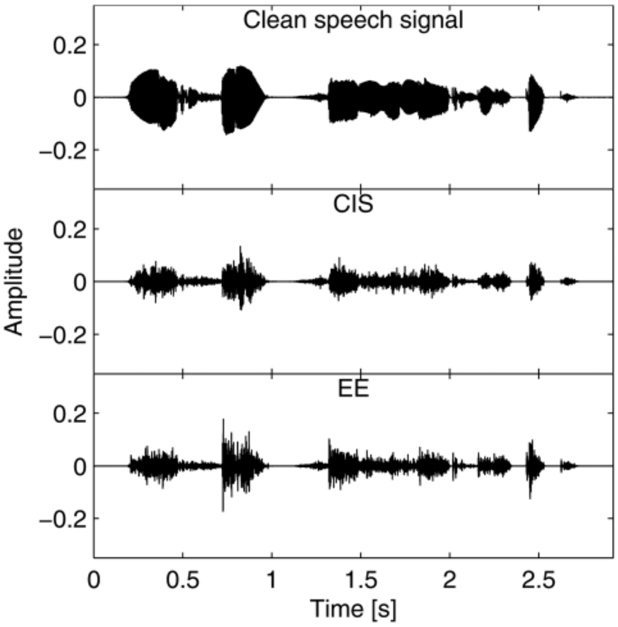

FIG. 2. Clean speech signal (above) and vocoded output (below) for both speech processing strategies CIS and EE of the Dutch sentence "Morgen gaan we naar de stad (Tomorrow we are going to the city)."

Analysis of the long-term spectrum of the speech signal showed that the differences in energy between the envelope enhanced speech and the CIS processed speech were primarily temporal. The long-term spectrum differed less than 0.3 dB per channel of the eight channel vocoder for the whole LIST speech material corpus.

The enhancement algorithm with emphasis on a possible real-time application was evaluated in experiment II. In this speech recognition task in stationary SSN, the peak signal was extracted from the noisy mixture of the speech and noise components of the signal with a Wiener filter algorithm in the front-end at the stage of the fast Fourier transform (FFT). The Wiener filter gain function $G(\lambda, n)$ is obtained as the minimum mean square error estimate of the complex spectral amplitude. The solution can be written as

$$G(\lambda, n) = \frac{|S(\lambda, n)|^2}{|S(\lambda, n)|^2 + |N(\lambda, n)|^2} = \frac{\xi(\lambda, n)}{1 + \xi(\lambda, n)}, \tag{3}$$

where S is the complex speech spectrum of the speech components and N of the noise components of the noisy speech signal Y. The instantaneous SNR is written as $\xi(\lambda, n)$ with the frame index λ and the frequency bin index n. The Wiener filter is an SNR dependent gain between zero and one that is applied in each time-frequency point. This gain is then applied to the noisy mixture $Y(\lambda, n) = S(\lambda, n) + N(\lambda, n)$ to obtain the noise reduced signal. A detailed derivation of the Wiener gain can be found in Vary and Martin (2006). The envelope of the noise reduced signal was afterwards obtained by a weighted sum of the power of the frequency bins like for the CIS strategy. This envelope was used as the input envelope E for the peak signal extraction algorithm shown in Fig. 1. The peak signal was finally added to the envelope of the noisy signal. Therefore, the processing is more similar to the processing that could be implemented in CIs. The nonlinearities that occur during the enhancement stage and the vocoding were also included, because the mixing of the noise and speech component was done before the vocoding of the signal.

Two different conditions that differed in the *a priori* knowledge about the signal and its components were evaluated with respect to the noisy mixture processed with the CIS strategy. In the condition that we refer to as EE_{IWF}, *a priori* knowledge of the noise and the speech components was used to calculate the instantaneous SNR in Eq. (3). The Wiener filter gain with a perfect estimate of the instantaneous SNR represents the ideal case in the front-end processing. In the second Wiener filter condition EE_{WF}, no *a priori* information was used to calculate the Wiener filter gain function. The latter represents a condition feasible in real-time implementation. To calculate the instantaneous SNR [Eq. (3)] the noise power estimation approach by Hendriks *et al.* (2008) was used.

In both conditions, after the channel representation of the noise reduced signal was obtained, it was used as the input envelope E in the peak extraction stage. The extracted peak signal was afterwards added to the noisy mixture. In comparison to the peak extraction in experiment I, the amplitude of the extracted peak signal was smaller due to the Wiener filter gain that weighted the noisy mixture with respect to the short term SNR. The mean peak value was 82% at 4 dB SNR and 73% at -4 dB SNR for the condition with the ideal Wiener filter in the front-end of the peak extraction stage as compared with the peak signal extracted from the clean speech.

The use of stationary SSN in speech for noise recognition tasks is not fully representative of a wider range of listening conditions. Therefore, the two talker condition was investigated in experiment III to generalize the outcomes to a broader range of masking stimuli and realistic listening situations. A female speaker was used as the target speaker, referred to as T, and the interfering speaker, referred to as I, was a male speaker. Three different enhanced conditions were evaluated with respect to the reference CIS processing in the two talker condition that differ in the input envelope to the peak extraction stage. In the first enhanced condition EE(T) + EE(I), the envelopes of the clean target speaker and the envelopes of the male interfering speaker were used separately as the input envelopes of the peak extraction stage. Therefore, amplified onsets were added for both speakers to the noisy envelope. The second condition EE(T + I) was obtained when using the noisy mixture as the input envelopes of the onset extraction stage. This condition represents the case that could be implemented without additional front-end algorithm and applied in noise environments. In the third enhanced condition EE(T) + I, only the onsets of the target speaker were enhanced. Hence, the input envelopes to the peak extraction stage were the envelopes of the target speaker. For consistency, the reference CIS condition where no additional enhancement is done is written as T + I.

C. Test materials and procedures

The sentences for the recognition tasks and the LR experiment were taken from the LIST (van Wieringen and Wouters, 2008). The female LIST sentences material consists of 35 lists of 10 Dutch/Flemish sentences spoken by a female speaker. Each sentence contains four to eight words and each lists consists of 32 or 33 keywords that were counted in the calculation of percent-correct scores. Each list is balanced to the phonetic distribution of conversational speech. The stationary SSN that was used in two of the three speech recognition experiments was obtained by taking the long-term averaged spectrum of all female LIST sentences.

The male speaker sentence corpus of the LIST sentences was used as the interfering speaker in the speech in speech recognition task. The male LIST sentences consist of 39 lists of 10 Dutch/Flemish sentences. Twenty of these 39 lists are unique in comparison to the female LIST sentences. The lists that contained sentences that were also included in the female LIST sentences were not used as the interfering speaker. Therefore, just speech materials that the subject was not acquainted with were used as the target and the interfering speaker.

In all experiments, the speech level was fixed at 65 dB sound pressure level (SPL) and the noise level was adapted to get the desired SNR. The subject had to repeat the sentence of the female target speaker.

The sentences were presented to the subjects at 5 SNRs from 4 to -4 dB with 2 dB steps in the stationary SSN in experiment I. Each subject listened to 230 sentences (= 30 training sentences + 10 sentences/condition × 5 SNR conditions × 2 speech processing strategies × test- and retest session). One list of ten sentences was used per strategy and per SNR. The 30 training sentences were presented at 4 dB SNR. Keyword percent scores were collected.

In experiment II, the sentences were presented to the subjects at the same SNR levels as in experiment I. Due to the fact that three different algorithms were tested, the total number of sentences increased to 330.

For the speech recognition experiment in the two talker situation, an adaptive procedure with a step size of 2 dB was used to determine the SRT where 50% of the keywords were understood correctly. Two sentences of the male speaker were randomly chosen from the speech material and concatenated with a break of 250 ms. The female speaker started randomly during the first sentence of the male speaker after 0.5 s up to 1.5 s and ended in the second running sentence of the male speaker. The SRT was calculated as the average of the SNRs of the last four responses. Each condition was presented four times across one session. With the training before the beginning of the task, each subject listened in total to 350 processed sentences across the two sessions.

For all experiments, a short training session with the lists that were not used in the test sessions was given to the subjects to get familiar with the noise vocoded stimuli and the respective task. During the training sessions, the answers were scored but they were excluded from the analysis of the results.

In the LR experiment, the first list of the LIST sentences was used in quiet. In each trial, the enhanced and the reference condition were presented to the subject. In total the subject listened to 20 sentences per trial. Each subject conducted the LR experiment three times on the same day of the retest of experiment I with breaks in between. The subject had to rate the loudness of the randomly selected stimulus on a scale from 0 (labeled with "SOFT") to 100 (labeled with "LOUD"). A loudness score of 50 was labeled with "OK." The subjects could replay unlimitedly the sentences before the input of the loudness score. All ten participating subjects performed the LR after the sentence recognition experiment. Hence, the same list could be used for all subjects without influencing the results.

In all experiments, all different conditions were randomly presented in each trial to avoid possible order effects and they were all performed double blind.

All test materials had a sample rate of 16 kHz and were digitized with a resolution of 16 bit. They were presented to the subject by using the software platform APEX 3 (Francart et al., 2008). A RME Multiface II DSP sound card was used to present the stimuli through a Sennheiser HDA 200 headphone. All tests took place in a double walled sound booth.

III. RESULTS

All percentage correct scores of the sentence recognition tasks were transformed to"rationalized" arcsine units as described in Studebaker (1985) for the statistical analysis with a repeated measures (RM) analysis of variance (ANOVA). For all experiments, a Bonferroni correction was used to correct the significance level of $p = 0.05$.

A. Experiment I

Group mean scores for the sentence recognition test in SSN with the ideal onset extraction and the mixing after the envelope enhancement are shown in Figs. 3 and 4. The percent correct scores for the CIS and EE strategies as a function of the SNR are shown in Fig. 3. The error bars depict the standard error of the mean.

A three-way RM-ANOVA was conducted with the factors strategy, SNR and session. Overall group mean scores for the EE strategy were significantly higher than those for the CIS strategy by 19% [$F(1,9) = 143.6$; $p < 0.001$]. An overall significant effect of SNR was also observed [$F(4,36) = 262.1$; $p < 0.001$]. The interaction between the factors strategy and SNR was significant [$F(4,36) = 9.6$; $p < 0.001$]. No significant effect of session or interaction between strategy and session were observed.

Post hoc analysis of the data at different SNRs showed that significant effects between the EE and the CIS strategy were obtained at signal-to-ratios of −4, −2, 0, and 2 dB which corresponded to a difference of 29.0% ($p < 0.001$), 30.8% ($p < 0.001$), 17.2% ($p < 0.05$), and 19.0% ($p < 0.05$), respectively.

Psychometric functions fitted to the data points are shown in Fig. 4. The variance of the SRT across the subjects is depicted by box plots. A cumulative Gaussian function was fitted to the percent correct scores at the five different SNRs and the SRT was determined for each strategy by use of the PSIGNIFIT MATLAB toolbox (Wichmann and Hill, 2001a, 2001b). A two-way RM-ANOVA with the factors strategy and session showed a significant effect of the strategy [$F(2,8) = 26.9$; $p < 0.001$] and no significant effect of the session. The SRT values for the CIS and EE strategies were −0.3 and −2.8 dB, respectively. Analysis of the pooled data showed that the SRT improvement of 2.5 dB by the EE strategy compared to the CIS strategy was significant [$F(1,19) = 48.9$; $p < 0.001$]. The analysis of the slope of the psychometric functions at the SRT showed a main effect of the strategy [$F(1,19) = 9.2$; $p < 0.05$]. The slope for the reference CIS strategy was 14.7.

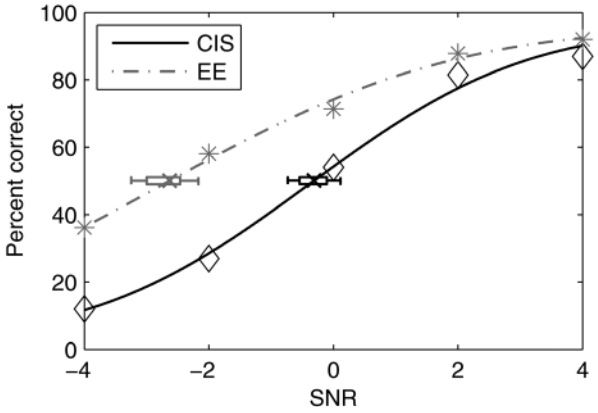

FIG. 4. Psychometric function fitted with a cumulative Gaussian function from the percent correct scores of experiment I as a function of the SNR in dB. The solid line represents the psychometric function obtained with the CIS and the dashed-dotted line with the EE strategy, respectively. Data points are marked with a diamond (CIS) and a star (EE). Box plots indicate the variance of the SRT from the psychometric functions of each subject.

B. Experiment II

The results of experiment II are shown in Figs. 5 and 6. The keyword percent correct scores for the three different conditions are shown as a function of the SNR in Fig. 5 with error bars depicting the standard error of the mean.

A three-way RM-ANOVA was conducted with the factors strategy, SNR and session. Significant effects were obtained for the factors strategy [$F(2,18) = 42.2$; $p < 0.001$] and SNR [$F(4,36) = 352.7$; $p < 0.001$]. The sessions were not significantly different. No other significant interaction

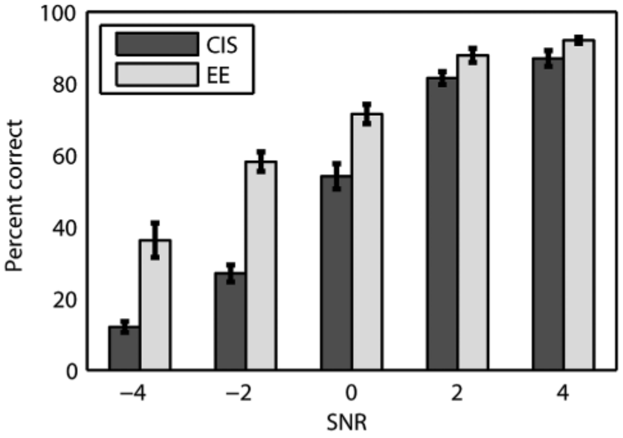

FIG. 3. Keyword understanding (percent correct) in experiment I for the strategies CIS and EE as a function of the SNR in dB. Error bars indicate the standard error of the mean.

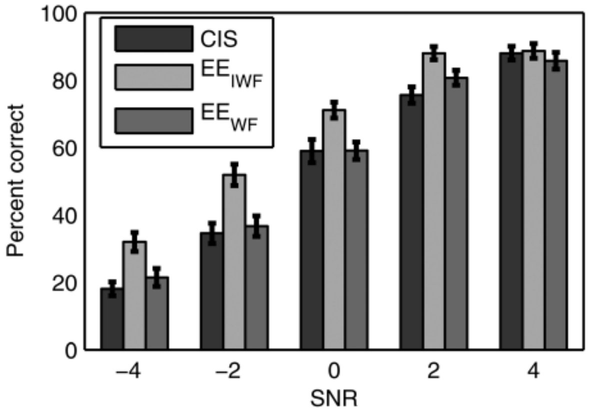

FIG. 5. Keyword understanding (percent correct) in experiment II for the strategies CIS, EE_{IWF} and EE_{WF} as a function of the SNR in dB. Error bars indicate the standard error of the mean.

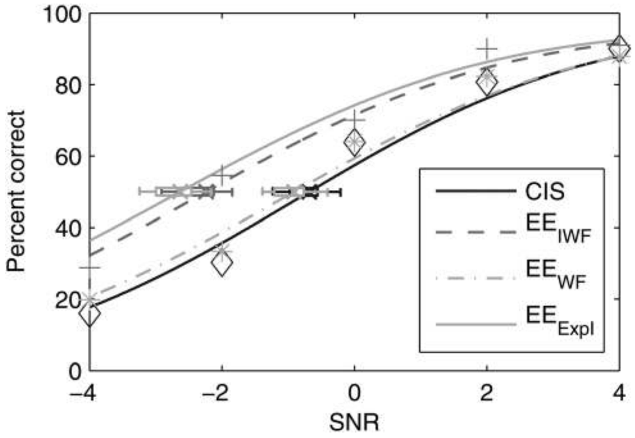

FIG. 6. Psychometric function fitted with a cumulative Gaussian function from the percent correct scores of experiment II as a function of the SNR in dB. The solid line represents the psychometric function obtained with the CIS, the dashed line with the EE_{IWF} and the dashed-dotted line with the EE_{WF} strategy, respectively. Data points are marked with a diamond (CIS), cross (EE_{IWF}) and a star (EE_{WF}). The psychometric function for the EE condition obtained in experiment I is plotted as the solid grey line. Boxplots indicate the variance of the SRT.

effects were obtained in experiment II. Therefore, the data was pooled for the subsequent analysis.

Post hoc analysis revealed that the difference between the CIS and the EE_{IWF} condition was overall significant ($p < 0.001$) which corresponded to an overall increase of 11.4% in keyword recognition. Also the increase of 9.6% in keyword recognition between the EE_{IWF} and the EE_{WF} condition was statistically significant ($p < 0.001$). The increase in speech intelligibility score of the EE_{IWF} with respect to the reference condition of 12.4% at 2 dB SNR, 12.2% at 0 dB SNR, 17.3% at −2 dB SNR, and 14% at −4 dB SNR was statistically significant. A comparison between the EE_{IWF} and the EE_{WF} condition revealed that the two conditions differed significantly at SNRs of 0, −2, and −4 dB which corresponded to an increase in speech intelligibility of 12.1%, 15.2%, and 10.6%, respectively. There were no significant differences between the CIS and the EE_{WF}.

In Fig. 6 the psychometric function derived from the percent correct scores of experiment II are plotted as a function of the SNR. The psychometric functions and the corresponding slope and SRTs were derived with the same procedure as in experiment I. Therefore, the boxplots represent the distribution of the SRT determined for each subject. For comparison purposes, the solid grey line represents the psychometric function obtained in experiment I for the EE condition. A two-way RM-ANOVA with the factors strategy and session showed a significant effect of the factor strategy [$F(2,18) = 33.6$; $p < 0.001$]. Like for the percentage correct scores, also no effect of the factor session was obtained. The SRT for the reference CIS condition was −0.7 dB SNR for the EE_{IWF} condition −2.4 dB SNR and for the EE_{WF} condition −0.9 dB SNR. The SRT improvements of 1.7 and 1.5 dB SNR of the EE_{IWF} with the CIS condition and the EE_{WF} condition were significant ($p < 0.001$). There was no significant difference obtained for the slope of the respective psychometric functions in a two-way RM-ANOVA. The slope was 12.7.

C. Experiment III

The pooled SRT values of both sessions obtained in experiment III for the four different conditions CIS, EE(T) + EE(I), EE(T + I), and EE(T) + I are shown in Fig. 7.

Analysis of the obtained SRT levels with a three-way RM-ANOVA with the factors session, strategy and trial revealed a significant effect of the factor strategy [$F(3,27) = 33.6$; $p < 0.001$] but neither an effect of the factor session nor trial. The mean SRT values for the four different conditions was 1.8 dB for the T + I, 1.7 dB for the EE(T) + EE(I), 1.9 dB for the EE(T + I) and −0.8 dB for the EE(T) + I condition. *Post hoc* analysis of the data revealed that the SRT improvement of 2.6 dB of the EE(T) + I in comparison with the T + I reference was significant ($p < 0.05$). Also the SRT improvements of 2.5 and 2.7 dB in comparison with the enhanced EE(T) + EE(I) and EE(T + I) condition, respectively, were significant ($p < 0.001$). There was no other significant difference in the SRT obtained in the pairwise comparisons of the other conditions.

D. Experiment IV: Loudness rating

The mean loudness score for the EE strategy was 53.2 and 52.9 for the reference CIS strategy on the scale of 0–100. The standard deviation was 8.2 for both strategies. A three-way RM-ANOVA was performed on the obtained data of the LR with the factors strategy, sentence and trial. No statistically significant effects of the strategy, sentence and trial were found. The data were averaged over the trials and the sentences for the subsequent analysis. The difference of the overall mean loudness score was only 0.3 and was not statistically significant.

IV. DISCUSSION

The envelope enhancement strategy EE was evaluated relative to the reference CIS speech processing strategy with eight channel noise vocoded speech in SSN and in a competitive talker situation. In total, three different sentence recognition tasks were performed to investigate the influence of the enhancement of the onsets in the speech envelope on

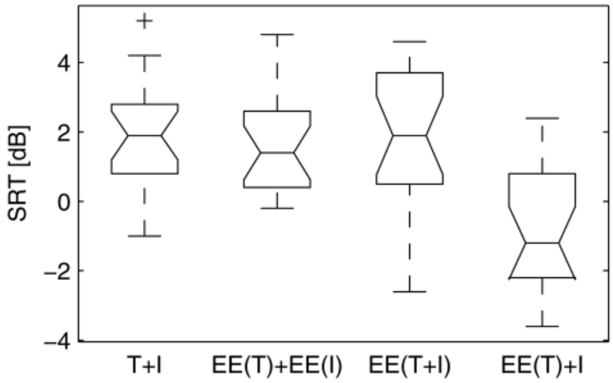

FIG. 7. Boxplots of the SRT obtained by an adaptive procedure of the test session for the four different conditions T + I (left), EE(T) + EE(I) (middle left), EE(T + I) (middle right), and EE(T) + I (right). The notches represent the confidence interval of the respective SRT value.

speech intelligibility and its applicability in real listening conditions.

Overall, the potential of increasing speech intelligibility by selectively enhancing the onset cues of the speech envelope provided by the EE strategy in comparison to the CIS strategy was shown in both noisy listening conditions. In general, the results demonstrate the importance of transient cues to speech intelligibility in CI vocoder simulations in both extreme cases of noisy or interfering background sounds such as stationary SSN and an interfering talker. The LR experiment analysis showed that there was no loudness difference in the perception of the CIS and the EE strategy. The amplified short-duration cues at the onsets of the speech envelope did not have an effect on overall loudness. Therefore, the improved speech perception was based on the envelope enhancement at the onsets in the speech envelope in each frequency band of the eight channel vocoder, because the onsets are the only cues affected by the processing while the stationary part remains the same for both strategies. This result is also underlined by the comparison of the long-term spectrum of the two strategies that did not differ much in the amplitude in each channel.

The significant improvement of the SRT of the envelope enhancement algorithms in experiments I, II, and III indicates that it was caused by the improved contrast between the speech and the noisy or interfering sounds, because the information of the boundaries of the phonemes is enhanced by amplifying the onsets of the speech envelope. We suggest that the enhanced representation of the onsets led to a segregation of the target speaker and the interfering background. Especially, when the onsets were emphasized in more than one frequency band simultaneously. This could have driven the attention of the listener to the target source (Shamma *et al.*, 2011). Although the rationale of the development of the EECIS strategy (Geurts *et al.*, 1999) was to include the bypassed rapid adaptation effect of the auditory nerve fibers in the signal processing of a CI, this approach led to an increased perception of the contrasts between the target and the interfering background in CI simulations with NH listeners with intact adaptation effect.

Results of the sentence recognition task in SSN are difficult to relate to other studies because to our knowledge no other study exists that investigated the effect of the enhancement of the onsets of a speech signal in vocoder simulations. The SRT obtained for the CIS strategy was about 2 dB lower than in the study by Dorman *et al.* (1998) for both mixing conditions that were vocoding the speech signal and the noise signal independently and adding them at the desired SNR afterwards or vocoding the noisy mixture. The difference in the SRT can be explained by the different speech material used in the respective experiments. The LIST sentences, developed for tests with severely hearing impaired listeners, have a lower speech rate in comparison to the H.I.N.T. sentences used in the study of Dorman *et al.* (1998) which might have affected the SRT obtained in both tasks.

van Wieringen and Wouters (2008) determined the SRT of the sentence speech tests used in this study for NH listeners and CI recipients. The SRT of the CI recipients was for the LIST sentences in a range of +0.5 dB for the best performers to +15 dB for the worst performers for the speech in noise test. In the second sentence recognition task in SSN, an SRT of −0.3 dB was obtained which overlaps well with the SRT of the best CI performers. Vocoder simulations for NH have been shown to be a good model and predictor to simulate parameter variations like the number of channels or intelligibility in noise when the processing is done similar to the processing in a CI (Dorman *et al.*, 1998; Friesen *et al.*, 2001). Even the SRT of the first sentence recognition task where the stationary SSN and the speech signal were vocoded independently from each other led to a SRT that was similar to the SRT of the best CI subjects with the same sentence material.

Chen and Loizou (2011) showed that the intelligibility of vocoded speech can be very well predicted when using coherence-based and speech transmission index based measures. To determine if the positive effect of the enhanced onsets of the envelope can also be predicted by an instrumental measure we tried to correlate the data of experiment I with the speech transmission based normalized covariance measure (Goldsworthy and Greenberg, 2004). The normalized covariance measure is, according to the speech transmission index, calculated as a weighted sum of transmission index values. In contrast to the speech transmission index, these values are based on the covariance between the reference signal and the processed output signal. For a detailed description the reader is referred to Goldsworthy and Greenberg (2004). The clean speech signal was used as the reference signal and the normalized covariance measure was determined for octave frequencies from 125 Hz to 8 kHz. While the measure shows a high correlation with the mean values for the reference CIS condition ($r = 0.98$; $p < 0.05$) and for the EE condition ($r = 0.99$; $p < 0.05$) it fails to predict that the EE scores are higher than the reference condition. The desired distortion of the envelope that led to the increased intelligibility is related to a lower value for the normalized covariance measure. The increase of speech intelligibility with speech intelligibility is not predicted with this objective performance measure for the type of envelope enhancement processing used in this study. The results show that the transmission index based measures are unable to predict speech intelligibility when nonlinear operations are involved in the processing (van Buuren *et al.*, 1999; Goldsworthy and Greenberg, 2004). Therefore, the other experiments are not evaluated with the normalized covariance measure.

To show the general feasibility of applying the envelope enhancement strategy on the noisy signal and actually extract the peak signal from the noisy signal, the second sentence recognition task in SSN was performed. The processing that was done in experiment II can similarly be implemented in CI processors.

In experiment II, the additional peak signal had a lower amplitude than in experiment I due to the SNR dependent weighting of the Wiener filter step [Eq. (3)]. Therefore, it is possible to compare the obtained percent correct scores of the two experiments with respect to the influence of the maximum gain on speech intelligibility. It is remarkable that the percent correct scores obtained with the EE strategy in experiment I and with the $\mathrm{EE_{IWF}}$ strategy are almost the same. The scores were slightly smaller at an SNR of −4 dB for the $\mathrm{EE_{IWF}}$ in comparison to the EE strategy in experiment I. This is most probably caused by the different amount of amplification that was provided at the onsets of the speech signal because if the mixing would be the cause of the difference in the scores it would most probably have also affected the scores at the other SNRs. In each channel of the vocoder the resulting peak signal has a smaller maximum peak value with the continuously weighting of the Wiener filter than in the case where the peak signal is extracted from the clean speech signal. The amplification factor A_{peak} remained the same in experiments I and II. With the same factor for A_{peak} in experiment II, the resulting total amount of amplification and its maximum value is smaller for the $\mathrm{EE_{IWF}}$ case. Therefore, the results suggest that the benefit is dependent on the maximum value of the amplification. Nevertheless, even the smaller amplification of the onsets of the speech envelope than in experiment I resulted in a speech intelligibility improvement. The fact that the improvement is largest for the maximum amplification suggests that the value should be chosen as high as possible or mapped to loud current levels in the CI application of the algorithm.

The results of the $\mathrm{EE_{WF}}$ strategy showed that there was no benefit of the envelope enhancement when a real noise power estimator is used for the Wiener filter front-end peak extraction at the SNRs that were tested in this study. As mentioned before, the speech material used in this study was developed for people with severe hearing loss. With speech material that is comparable to continuous discourse, the SRT of CI users is considerably higher (Hu and Loizou, 2010; van Wieringen and Wouters, 2008). However, our results showed no detrimental effects of the onset enhancement either, because no decrease in speech intelligibility at the different SNRs was observed. The resulting processed signal is robust to misdetections and providing a non-ideal amplification did not lead to a decrease in speech intelligibility.

To quantify the difference in the amount of amplification of the added onset peaks between experiment I and for both Wiener filter conditions in experiment II, a comparison of the onset peak amplitude between the $\mathrm{EE_{IWF}}$ and the $\mathrm{EE_{WF}}$ case is shown in Fig. 8 across all 35 processed lists of the LIST sentences. The peak amplitude was calculated by determining the maximum value of each peak in the additional peak signal in each frequency band. The averaged amplitude over all onset peaks is normalized to the amplitude that is provided by the EE algorithm when the enhancement of the onset cues is processed based on the clean speech signal. The normalized amplitude is shown for the SNR range that was tested in this study up to 15 dB. The maximum variation across the different normalized peak amplitudes for the LIST sentences considered along the simulated SNRs is about 2 % and 4 % for the $\mathrm{EE_{IWF}}$ and $\mathrm{EE_{WF}}$ processing, respectively. It is shown that the normalized amplitude did not differ much across all lists of the speech material at the SNRs used in this study for both envelope enhancement strategies. The amplification by $\mathrm{EE_{IWF}}$ is close to the value obtained for the enhancement of the clean speech signal and still about 70% at an SNR of −4 dB. In contrast, for $\mathrm{EE_{WF}}$ the cumulative amplitude is much smaller. At SNRs below 10 dB, the performance of the Wiener filter on the mixed signal decreases and more onsets are missed due to a drop in performance of the noise power estimator. Especially at high SNRs where the detection of the onsets is fairly reliable for $\mathrm{EE_{WF}}$, a higher gain factor A_{peak} could be chosen to provide the same amount of amplification as in the EE for the clean speech signal or the $\mathrm{EE_{IWF}}$ case, respectively. A simulation of the peak extraction revealed that at an SNR of around 3 dB there are still 75% of the peaks detected with a false alarm rate of 16%. Therefore providing a higher gain factor A_{peak} should be able to provide the same amount of amplification at least in the region above 3 dB without adding too much distortion due to a wrong peak detection to the mixed signal. The operating SNR range in the two experiments in stationary SSN was from 4 to −4 dB where the $\mathrm{EE_{WF}}$ leads to an amplification lower than 0.4 with respect to the amplification in experiment I which is partly caused by a failure in the detection and also by providing a lower amplitude to the correct detected onsets. This analysis suggests that the amplification that was chosen too small for the $\mathrm{EE_{WF}}$ processing and the factor A_{peak} should be increased.

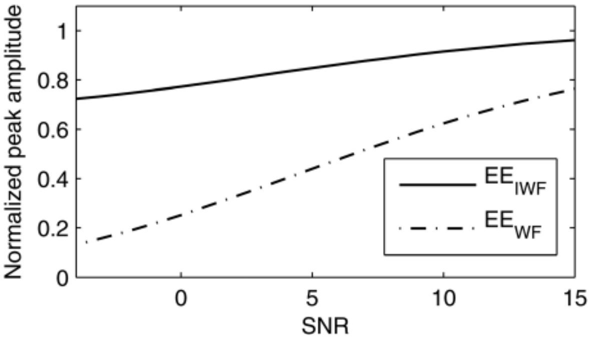

FIG. 8. Amplitude of the onset peaks of all lists of the LIST sentences for the signal processing strategies $\mathrm{EE_{IWF}}$ (solid line) and $\mathrm{EE_{WF}}$ (dash-dotted line) normalized to the onset peaks of the EE processing for the clean speech signal (as in experiment I) as a function of the SNR.

Figure 9 shows the correlation between the mean scores obtained at the tested SNRs values and the normalized amplitude of the peak signal for the $\mathrm{EE_{IWF}}$ and the $\mathrm{EE_{WF}}$ signal processing. To investigate the relationship between the peak amplitude and the percent correct scores, a multiple regression analysis was performed for both strategies with the factors amplitude and SNR. Individual data points for all subjects in both sessions are marked with a circle and a cross for the $\mathrm{EE_{IWF}}$ and the $\mathrm{EE_{WF}}$ strategy, respectively. The removal of outliers that differed more than two standard deviations from the obtained regression, resulted in 97 and 92 data points for the $\mathrm{EE_{IWF}}$ and the $\mathrm{EE_{WF}}$ processing, respectively. The correlation between the normalized amplitude and the keyword percent correct scores was in both processing cases statistically significant. Using the multiple regression, the fitted model for the $\mathrm{EE_{IWF}}$ processing was significant [$F(2,95) = 140.9$; $p < 0.001$; adjusted $R^2 = 0.75$] with the significant standardized coefficient of $\beta_{\mathrm{SNR}} = 0.94$ ($p < 0.001$). The factor amplitude was not significant. For the $\mathrm{EE_{WF}}$ processing, also a significant model

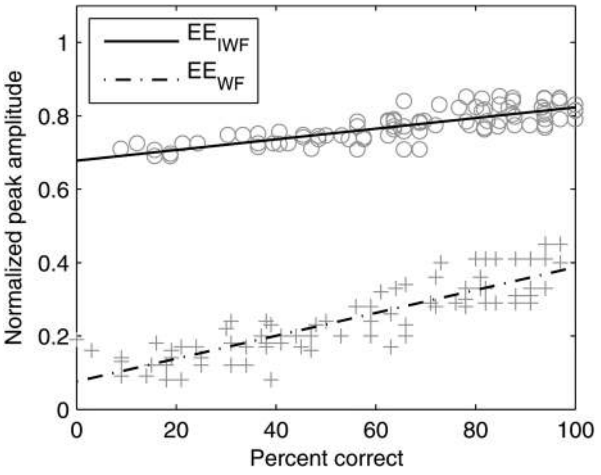

FIG. 9. Correlation between the keyword percent correct scores and the normalized amplitude for the EE$_{\mathrm{IWF}}$ (solid line) and the EE$_{\mathrm{WF}}$ (dash-dotted line) case. The circles mark the data points obtained with the EE$_{\mathrm{IWF}}$ strategy while the crosses represent the data points obtained with the EE$_{\mathrm{WF}}$ strategy for all subjects in both sessions.

emerged [$F(2,90) = 207$; $p < 0.001$; adjusted $R^2 = 0.82$]. The predictor factors SNR [$\beta_{\mathrm{SNR}} = 0.68\,(p < 0.001)$] and amplitude [$\beta_{\mathrm{amp}} = 0.24 \ \ (p < 0.05)$] were significant. The fitted linear regression lines are shown as well in Fig. 9 with the solid line representing the EE$_{\mathrm{IWF}}$ and the dash-dotted line the EE$_{\mathrm{WF}}$ signal processing. This result suggests that the value of the amplification of the onset peaks is the driver for the higher speech intelligibility performance. As mentioned before, providing a combination of a better peak extraction and a higher overall gain factor A_{peak} may result in an increased performance as observed for EE in experiment I or for EE$_{\mathrm{IWF}}$.

The results of the comparison show that in the target SRT\SNR range for CI recipients it is possible to detect the onsets and process the envelope enhancement on the noisy signal. Taking the results of the EE$_{\mathrm{WF}}$ scores into account, no detrimental effects have to be expected when missing some of the onsets or amplifying wrong parts of the noisy signal. This suggests that there is a margin for errors in the peak extraction. Note however that the algorithm is not optimized for combination with a noise reduction algorithm in the front-end. The onsets in the envelope could, for example, be derived from an estimate of the speech power or the instantaneous SNR. The comparison between the input envelope and the amplified slow envelope in Fig. 1 is sensitive to a residual noise floor.

Stationary SSN is not always representative of the background sounds in real adverse listening environments and represents one of the extreme cases of noisy maskers. Another challenging situation where especially HI listeners and CI users have grave difficulties is when another talker is interfering with the target speaker. The result shown in Fig. 7 of the sentence recognition task with an interfering speaker demonstrated a significant SRT improvement for the condition when just the target speaker is enhanced [EE(T) + I]. This task showed that the enhanced onset cues of the target speaker signal can also increase speech intelligibility in listening conditions with non-stationary noise maskers.

It is interesting to point out that for both other enhancement conditions no decrease in speech intelligibility was apparent in comparison to the standard CIS strategy. This suggests that an SRT improvement can only be obtained by amplifying reliably the onset cues of the target speaker. Comparing the results of the EE(T) + I with the results of the EE(T + I) and EE(T) + EE(I) conditions, there is a negative effect of also enhancing onsets of the interfering speaker signal. But the SRT is not higher in any enhanced condition than in the reference CIS condition.

In both target-interfering sound scenarios, we suggest using a noise reduction algorithm in the front-end for the extraction of the onsets of the target signal. It is possible to apply the noise reduction step to the noisy speech signal and additionally amplifying the onsets of the resulting noise reduced signal in the envelope extraction stage. It is not clear if the enhancement of the onsets would lead to an additional benefit in terms of speech intelligibility in comparison to the noise reduced signal. This should be investigated further. A major advantage of our approach to use the noise reduction algorithm in the front-end just for the peak signal extraction is that no artifacts like speech distortions or musical noise that occur with state-of-the-art noise reduction algorithms are introduced. All speech information is present in the signal presented to the subject. Therefore, the detrimental effects of estimation errors possibly leading to reduced speech intelligibility are avoided. The results suggest that adding the extracted signal to the noisy envelope even with misdetections does not decrease speech intelligibility in comparison to the reference CIS strategy. But the primary aim of the study was to investigate the effect of enhanced onset cues on speech intelligibility in different interfering background sounds. The use of a front-end noise reduction step to extract the onsets of the target signal from the noisy speech signal (experiment II) supports the real-time applicability of the approach at least for SSN noise.

Learning effects play often a role in studies with NH listeners and CI simulations (Fu and Shannon, 1999; Davis *et al.*, 2005). Therefore, all the tests were conducted in a test-retest design with several days between the two sessions. In all three experiments, no statistically significant session effect was obtained. The ideal enhanced conditions provided an immediate significant improvement without the need of training and feedback.

It is not clear if the full potential of the onset enhancement strategy is achieved because the CV increase applied here is the same for all subjects. Kennedy *et al.* (1998) showed that the increase must be individually adjusted for each HI listeners to obtain the optimal performance in a consonant recognition task. Individualized amplification factors could have led to better scores. But this effect should have been relatively small here, because only NH subjects participated in our experiments. In the study of Kennedy *et al.* (1998), it was not reported if the increase was also manually adjusted for NH listeners to reach the optimal performance level. Furthermore, Hazan and Simpson (1998) showed that treating all different phonemes in the same way is not the best option to achieve maximum speech intelligibility. Hence, including an amplification that is optimized for the different phonemes of speech could improve speech intelligibility even more. Overall, our results suggest that a higher amplification leads to more benefits in terms of speech intelligibility.

It is difficult to compare the results obtained in the study with other studies that were done with CI users. While the TESM strategy showed a significant improvement of 11.3% at an SNR of 5 dB in multitalker babble noise (Vandali, 2001), the same strategy did not differ significantly in a sentence recognition task in Holden *et al.* (2005) when the target signal level was presented at 65 dB SPL. Bhattacharya *et al.* (2011) even observed a decrease in speech intelligibility in quiet for the strategy and obtained a significant improvement when adding an additional spectral extension stage to the processing. The TESM strategy is more focused on the transient part and not the onset of the speech envelope. Therefore, it is hard to compare the results obtained in our sentence recognition task with the results obtained with the TESM strategy in CI users.

The proposed EE algorithm is based on the EECIS that was developed by Geurts *et al.* (1999) for CI. The EECIS algorithm was only tested in quiet with CI recipients and on the word level with CVC words and stop consonants. Small non significant differences between the enhancement algorithm and the reference strategy were obtained. Taking the results of this study into account, these results were most probably not significantly different due to ceiling effects that could also explain the results in this study at high SNRs. The EE and the EECIS strategy both have the advantage that they introduce no additional time delay in the processing chain of the CI. Stone and Moore (2005) have shown that a processing delay up to 20 ms is tolerable in all CI users to avoid a disturbing asynchrony between lip-reading and sound perception. This time can be used for additional processing like a noise reduction system. The total additional time delay to the time required for the reference CIS processing is, e.g., determined by the delay of the estimator to calculate the weighting function of the Wiener filter in experiment II.

V. CONCLUSION

In summary, this study demonstrated the importance of onset cues for speech intelligibility in noisy listening conditions. The effects of the onset enhancement strategy EE in SSN and in the competitive talker condition was investigated with noise vocoded speech. A significant improvement in speech intelligibility was obtained for the conditions when the peak extraction was done under the assumption of *a priori* knowledge of the clean speech signal. It is possible to implement the EE algorithm in real-time and apply it to the noisy speech signal in stationary SSN. The proposed signal processing approach may be similarly applicable to hearing aids and CIs. The algorithm is not effective in real-life situations at low SNR, even with state-of-the-art single channel noise reduction strategies in the front-end of the processing. However, the findings suggest that speech enhancement strategies for HI and CI listeners could lead to a significant benefit when the onsets of the speech envelope of the target are enhanced. The effect on speech intelligibility in adverse listening conditions by applying the processing in hearing aids and CIs should be investigated further.

ACKNOWLEDGMENTS

R.K. was supported by the EU within the Marie Curie ITN AUDIS, Grant Agreement No. PITN-GA-2008-214699. We thank Jolijn Monstrey and Nele Meybosch for significant help in accumulating the data and are very thankful to all subjects who voluntarily participated in our experiments.

Apoux, F., Tribut, N., Debruille, X., and Lorenzi, C. (**2004**). "Identification of envelope-expanded sentences in normal-hearing and hearing-impaired listeners," Hearing Res. **189**, 13–24.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.

Benesty, J., Makino, S., and Chen, J. (**2005**). *Speech Enhancement*, 1st ed. (Springer, Berlin), pp. 9–66.

Bhattacharya, A., Vandali, A. E., and Zeng, F.-G. (**2011**). "Combined spectral and temporal enhancement to improve cochlear-implant speech perception," J. Acoust. Soc. Am. **130**, 2951–2960.

Boll, S. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans Audio Speech Lang. Proc. **27**, 113–120.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA).

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. Acust. **86**, 117–128.

Chen, F., and Loizou, P. C. (**2010**). "Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing," Ear Hear. **31**, 259–267.

Chen, F., and Loizou, P. C. (**2011**). "Predicting the Intelligibility of Vocoded Speech," Ear Hear. **32**, 331–338.

Chen, F., and Loizou, P. C. (**2012**). "Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise," J. Acoust. Soc. Am. **131**, 4104–4113.

Chen, J., Benesty, J., Huang, Y., and Doclo, S. (**2006**). "New insights into the noise reduction Wiener filter," IEEE Trans Audio Speech Lang. Proc. **14**, 1218–1234.

Clarkson, P. M., and Bahgat, S. F. (**1991**). "Envelope expansion methods for speech enhancement," J. Acoust. Soc. Am. **89**, 1378–1382.

Davis, M. H., Johnsrude, I. S., Hervaix-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241.

Delgutte, B., and Kiang, N. Y. S. (**1984**). "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," J. Acoust. Soc. Am. **75**, 897–907.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.

Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.

Francart, T., van Wieringen, A., and Wouters, J. (**2008**). "APEX3: A multi-purpose test platform for auditory psychophysical experiments," J. Neurosci. Meth. **172**, 283–293.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Fu, Q.-J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded speech in noise with nonlinear amplitude mapping," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. 369–372.

Geurts, L., and Wouters, J. (**1999**). "Enhancing the speech envelope of continuous interleaved sampling processors for cochlear implants," J. Acoust. Soc. Am. **105**, 2476–2484.

Goldsworthy, R. L., and Greenberg, J. E. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **119**, 1727–1739.

Hazan, V., and Simpson, A. (**1998**). "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Commun. **24**, 211–226.

Hendriks, R. C., Jensen, J., and Heusdens, R. (**2008**). "Noise tracking using DFT domain subspace decompositions," IEEE Trans Audio Speech Lang. Proc. **16**, 541–553.

Hendriks, R. C., and Martin, R. (**2007**). "MAP estimators for speech enhancement under normal and rayleigh inverse gaussian distributions," IEEE Trans Audio Speech Lang. Proc. **15**, 918–927.

Holden, L. K., Skinner, M. W., Fourakis, M. S., and Holden, T. A. (**2005**). "Speech recognition with the advanced combination encoder and transient emphasis spectral maxima strategies in nucleus 24 recipients," J. Speech. Hear. Res. **48**, 681–701.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Hu, Y., and Loizou, P. C. (**2007a**). "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun. **49**, 588–601.

Hu, Y., and Loizou, P. C. (**2007b**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.

Hu, Y., and Loizou, P. C. (**2010**). "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," J. Acoust. Soc. Am. **127**, 3689–3695.

Hu, G., and Wang, D. (**2007**). "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio Speech Lang. Proc. **15**, 396–405.

Kennedy, E., Levitt, H., Neuman, A. C., and Weiss, M. (**1998**). "Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners," J. Acoust. Soc. Am. **103**, 1098–1114.

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (**2007**). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 2365–2375.

Kluender, K. R., Coady, J. A., and Kiefte, M. (**2003**). "Sensitivity to change in perception of speech," Speech Commun. **41**, 59–69.

Langhans, T., and Strube, H. W. (**1982**). "Speech enhancement by nonlinear multiband envelope filtering," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. 156–159.

Lee, J. H., and Kewley-Port, D. (**2009**). "Intelligibility of interrupted sentences at subsegmental levels in young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 1153–1163.

Lewicki, M. S. (**2010**). "A signal take on speech," Nature **466**, 821–822.

Lorenzi, C., Berthommier, F., Apoux, F., and Bacri, N. (**1999**). "Effects of envelope expansion on speech recognition," Hear. Res. **136**, 131–138.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., and Spriet, A. (**2010**). "Multicenter evaluation of signal enhancement algorithms for hearing aids," J. Acoust. Soc. Am. **127**, 1491–1505.

Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **121**, 1709–1716.

Owren, M. J., and Cardillo, G. C. (**2006**). "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," J. Acoust. Soc. Am. **119**, 1727–1739.

Plomp, R. (**1988**). "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," J. Acoust. Soc. Am. **83**, 2322–2327.

Rasetshwane, D. M., Boston, J., Durrant, J. D., Li, C.-C., and Genna, G. (**2009**). "Enhancement of speech intelligibility using transients extracted by wavelet packets," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 173–176.

Schroeder, M. R. (**1981**). "Modulation transfer functions: Definition and measurement," Acustica **49**, 179–182.

Shamma, S. A., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Skowronski, M. D., and Harris, J. G. (**2006**). "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environment," Speech Commun. **48**, 549–558.

Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. USA **107**, 12387–12392.

Stone, M. A., and Moore, B. C. J. (**2005**). "Tolerable hearing-aid delays: IV. Effects on subjective disturbance during speech production by hearing-impaired subjects," Ear Hear. **26**, 225–235.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

van Buuren, R. A., Feesten, J. M., and Houtgast, T. (**1999**). "Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality," J. Acoust. Soc. Am. **105**, 2903–2913.

Vandali, A. E. (**2001**). "Emphasis of short-duration acoustic speech cues for cochlear implant users," J. Acoust. Soc. Am. **109**, 2049–2061.

van Wieringen, A., and Wouters, L. (**2008**). "LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands," Int. J. Audiol. **47**, 348–355.

Vary, P., and Martin, R. (**2006**). *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, 1st ed. (Wiley & Sons, Ltd., Chichester), pp. 389–466.

Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671–1684.

Wichmann, F. A., and Hill, N. J. (**2001a**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.

Wichmann, F. A., and Hill, N. J. (**2001b**). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," Percept. Psychophys. **63**, 1314–1329.

Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C.-C., Durrant, J. D., Kovacyk, K., and Shaiman, S. (**2007**). "Speech signal modification to increase intelligibility in noisy environments," J. Acoust. Soc. Am. **122**, 1138–1149.