Methodological aspects of an adaptive multidirectional pattern search to optimize speech perception using three hearing-aid algorithms

Franck, B.A.M.; Dreschler, W.A.; Lyzenga, J.

DOI
10.1121/1.1808220

Publication date
2004

Published in
The Journal of the Acoustical Society of America

Methodological aspects of an adaptive multidirectional pattern search to optimize speech perception using three hearing-aid algorithms

Bas A. M. Franck[a)] and Wouter A. Dreschler
Academic Medical Center, Clinical and Experimental Audiology, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

Johannes Lyzenga
Department of Otolaryngology, VU Medical Center, De Boelelaan 1117, 1081 HV, Amsterdam, The Netherlands

(Received 8 October 2003; revised 19 August 2004; accepted 23 August 2004)

In this study we investigated the reliability and convergence characteristics of an adaptive multidirectional pattern search procedure, relative to a nonadaptive multidirectional pattern search procedure. The procedure was designed to optimize three speech-processing strategies. These comprise noise reduction, spectral enhancement, and spectral lift. The search is based on a paired-comparison paradigm, in which subjects evaluated the listening comfort of speech-in-noise fragments. The procedural and nonprocedural factors that influence the reliability and convergence of the procedure are studied using various test conditions. The test conditions combine different tests, initial settings, background noise types, and step size configurations. Seven normal hearing subjects participated in this study. The results indicate that the reliability of the optimization strategy may benefit from the use of an adaptive step size. Decreasing the step size increases accuracy, while increasing the step size can be beneficial to create clear perceptual differences in the comparisons. The reliability also depends on starting point, stop criterion, step size constraints, background noise, algorithms used, as well as the presence of drifting cues and suboptimal settings. There appears to be a trade-off between reliability and convergence, i.e., when the step size is enlarged the reliability improves, but the convergence deteriorates.
[DOI: 10.1121/1.1808220]

PACS numbers: 43.66.Yw, 43.66.Ts [DDO] Pages: 3620–3628

I. INTRODUCTION

In previous studies, multi-directional strategies have been proposed to optimize hearing-aid parameters (e.g., Neuman *et al.*, 1987; Kuk and Lau, 1996). In many studies the (modified) Simplex procedure has been promoted. The Simplex procedure is especially advantageous over tournaments with respect to time efficiency. Moreover, the procedure can take into account interactions between parameters by considering multiple responses at the same time. However, many methodological consequences have not been studied systematically.

In order to apply the Simplex procedure in the auditory domain, a paired-comparison paradigm has been used most frequently.[1] There are several reasons for using paired-comparisons. It often appears problematic to use objective scores like speech intelligibility thresholds, because different parameter settings do not always yield different scores. Other important advantages of the paired-comparison paradigm comprise high sensitivity and psychological advantage due to a high motivation by active subject interaction. For audiological purposes it appears especially advantageous to evaluate not only speech intelligibility, but also listening comfort in order to guarantee better satisfaction from using the hearing aid.

In a pilot study we expanded the domain of the modified Simplex procedure of Neuman *et al.* (1987) from two dimensions to three. The dimensions represented different algorithms instead of different parameters of a single algorithm. The results indicated that the test-retest reliability of the optimization procedure is relatively poor due to the presence of settings that were hard to distinguish. Moreover, listening comfort appeared a more reliable evaluation criterion than speech intelligibility. Therefore, in this study we selected the listening comfort criterion to determine subjective preferences.

In order to improve the reliability we modified the procedure in the following ways. First, the step size in the procedure has been made adaptive. That is, the step size for each dimension was governed by the perceptual differences between settings for each listener. Initially, the step size was large. After three rounds of comparisons the step size was halved. The process of halving continued until the listener could not distinguish the settings any more. In that case the step size was enlarged for that particular dimension. Second, boundary effects for the constrained complex of settings have been made less influential. This was achieved on the one hand by extreme algorithm settings at the boundary that

[a)]Author to whom correspondence should be addressed. Clinical and Experimental Audiology, Academic Medical Center Amsterdam, P.O. Box 22660, 1100 DD, Amsterdam, The Netherlands. Telephone: +31-20-5663918; Fax: +31-20-5669068. Electronic mail: b.a.franck@amc.uva.nl

are assumed to be unfavorable. On the other hand the number of inside settings relative to boundary settings was increased. This was achieved by increasing the number of settings from five to ten for each dimension. We hypothesize an increased reliability of the procedure as a consequence of these modifications.

Although the necessity of distinguishable settings is assumed to be an important prerequisite for a reliable optimization procedure, other procedural and nonprocedural factors can be of influence as well. Procedural factors that need closer inspection are the initial setting, the stop criterion, and the maximum number of paired comparisons. This latter constraint prevents listeners from losing their attention. When different initial settings yield different final settings (optima), the optimization procedure performs inconsistently. Next to this effect, the reliability might depend on the initial setting for reasons yet unknown. The stop criterion is an important tool for proper convergence. When the stop criterion has been chosen falsely, the procedure might end too quickly. The procedure can also end too late. This can happen when the, a priori defined, maximum number of paired comparisons has been reached. In both cases at best a sub-optimal setting will be found.

Non-procedural factors that can influence the reliability in a negative way can be caused by acoustical constraints and cognitive factors. When the stimulus material has been recorded real-life, a smearing effect can be introduced as a result of reverberation. A shaping effect in the frequency domain can be expected when the stimuli are presented through a hearing aid. Both effects could have detrimental effects on the reliability. Although optimization of speech perception for hearing aids should be performed in realistic acoustical conditions in future studies, in this study we used artificial speech and noises that are presented by headphones. So, both a stationary car noise and a fluctuating speech-shaped noise have been chosen as experimental parameters. The most important cognitive factor is the criterion that is used by the listener to judge which fragment is better. That is, the listener's criterion can drift during an experiment. This aspect needs close inspection.

In this study three experimental questions will be answered. Does the adaptive multidirectional pattern search perform reliably? Or, stated differently, does the adaptive step size result in high reliability of the optimization strategy? Directly related to this, what are the effects of procedural and nonprocedural factors on the reliability of the multidirectional pattern search? Finally, taking all considerations into account, we evaluate how applicable the multidirectional pattern search is to the auditory domain for the selection of an optimal setting of complex hearing aids.

II. METHOD

A. Algorithms

Three experimental auditory signal-processing algorithms are used in the optimization procedure. The individual dimensions are spectral enhancement (SE), spectral lift (SL), and noise reduction (NR), developed and described in detail by Lyzenga *et al.* (2002).

The spectral enhancement algorithm expands the spectral peaks for frequencies between 200 and 5000 Hz. In that region slow spectral fluctuations are slightly compressed, intermediate fluctuations expanded, and fast fluctuations remain intact to avoid statistical variations of noise from being expanded. The spectral enhancement was realized using a power function. Enhancement factor M represents the exponent, the spectral amplitude of the signal a represents the base of the power function.

The spectral lift algorithm is a linear filter that is assumed to reduce upward-spread-of-masking. That is, to try prevent low-frequency high-level formants from masking higher-frequency low-level formants. Therefore, the algorithm is designed to progressively emphasize frequencies between 0.3 and 3 kHz, the region in which the second to fourth formant are dominant. For frequencies higher than 3 kHz the spectral envelope decreases again, so the filter is sawtooth shaped, and the peak of the filter is centered on 3 kHz. Emphasis of high frequencies is avoided, because sharp sounds are assumed to be uncomfortable.

The noise reduction algorithm attempts to suppress noise based on the temporal behavior of the signal, i.e., the phase variations. When the temporal behavior of spectral components is irregular, the signal is assumed to be noisy, and the linear suppression filter is given small amplitudes. The amplitude of the filter is set high when spectral components behave temporally regular, and are assumed to be speech-like signals.

It was found to be difficult to assemble a series of ten easily discernable settings for each algorithm without introducing intolerable distortions. To overcome this difficulty, we included settings that, at first inspection, seemed counter-intuitive. Not only spectral expansion is allowed but also spectral compression, not only spectral lift but also spectral suppression, and finally, not only noise reduction but also noise amplification. The settings are labeled from -3 to 6, such that 0 is the reference setting.

The spectral enhancement factor M ranges between 0.25 and 2.5 in steps of 0.25 dB/dB. For setting 0, $M=1$ and there is no spectral enhancement. The spectral lift settings have values between -9 and 18, in nine 3 dB steps. A lift-value of 0 represents no spectral lift. The noise reduction multiplication factors are represented by dimensionless values between -3 and 6 in steps of 1, and setting 0 corresponds to a value of 0. The algorithm values have been chosen so that steps are approximately perceptually equidistant.

B. Experimental procedures and conditions

1. Multidirectional pattern search with adaptive step size

In order to simultaneously optimize multiple hearing aid algorithms a multidirectional pattern search design is used in this study. We use three dimensions that represent three hearing aid algorithms. Several procedural features characterize the search strategy, which is aimed at finding the optimal combination of all algorithm settings. The search has a single initial setting. After the first comparison, the next setting is assumed to have improved performance, i.e., listening comfort in this study. To search for improved performance, the setting under focus is compared pair wise with three neighbors (n paired comparisons for n dimensions). The settings that are included for paired comparisons are orthogonally related to the initial setting under focus and remain orthogonally related. So, three orthogonal search directions are used throughout the procedure for this pattern search design. The decision as to what will be the next setting under focus depends on the outcome of all three paired comparisons in the following way. When neither of the neighbors wins, the setting under focus remains the best setting. When one neighbor wins, this neighbor will be the new best setting. When two or three neighbors win, the new setting is estimated from the winning directions. That is, the *assumed* new best setting is found by addition of the vectors that represent the search directions, i.e., the directions of paired comparisons. As a consequence, the progress can be along a diagonal with respect to the orthogonal system of search directions.

The new search directions for the novel best setting depend on the history of search directions. For each dimension the search direction remains the same when the neighbor yields higher listening comfort. The search direction is reversed when the setting at stake performs better than the neighbor in that particular direction. The incorporation of the history of search directions is very important for this search strategy in order to accelerate the optimization process. The settings can never escape a predetermined lattice structure. This feature is necessary for our study due to the fact that we had to record the speech and noise material in advance.

The step size has been made adaptive, and can be characterized by perceptual distinctness. At the start of the search, the step size is equal to four steps of the lattice structure. After one series of three paired comparisons the step size is halved to two steps. After another three rounds the step size is halved to one step, the minimum step size. The halving process has been incorporated to speed convergence. The initial large step size was mainly chosen to ensure audible differences between settings and to make the listeners acquainted with the perceptual dimensions. At any time the fragments to be compared can be repeated. When the listener cannot distinguish between the two settings, the step size for that particular dimension will be increased until it is equal to three. Again, after three series the step size will be halved and rounded, until the step size is equal to one step. It is important to note that the ‘‘no difference’’ button can be used only after the ‘‘repeat’’ button has been used once, as the listeners should be convinced that the fragments are indistinguishable, instead of just rather difficult to judge in terms of listening comfort. Each pair of fragments can be repeated once.

The multidirectional pattern search procedure terminates when one of the following three criteria is met. When the same setting is frequented three times or when a setting wins a paired comparison more than six times, the procedure will stop. These stop criteria are a compromise between good convergence and limited required average test time. The third stop criterion is the maximum number of paired comparisons in the procedure. This criterion has been included to limit the maximum test time of the procedure. In a pilot experiment a limit of 54 paired comparisons was found to be a reasonable choice, which corresponds to a test duration of fifteen minutes on average.

For four different starting points a test (T) and a retest (R) have been performed. In the multidimensional setup we used combinations $T(-3)=R(-3)=(-3,-3,-3)$, $T(1)=R(1)=(1,1,1)$, $T(3)=R(3)=(3,3,3)$, and the center of gravity of the winners of $T(-3)$, $R(-3)$, $T(1)$, $R(1)$, $T(3)$, and $R(3)$. This starting point, which could be a suboptimal setting instead of an optimal setting is labeled as T(S) or R(S) depending on the test or retest status.

2. Multidirectional pattern search with step size as parameter

The experimental procedure for this experiment is the same as for experiment 1, except for the step size. Using the step size as a parameter, the dependence of the reliability on the step size can be estimated. In this experiment, the step size is either fixed or variable. The fixed step size is either 3 or 4 lattice grid steps. The variable step size has a lower limit of 2 and an upper limit of 3 steps. Therefore, only the lower limit is different from the lower limit of experiment 1, which was 1 step. Hence, in combination with the first experiment there are four step size configurations. The labeling of the configurations follows the possible step sizes in the procedure. So, the procedure of the experiment 1 is step123, the variable step size with elevated minimum step size is labeled step23. The fixed step size configurations are labeled step3 and step4. The first ‘‘orienting’’ step of the multidirectional pattern search is twice as large as the fixed step size throughout the procedure, i.e., 6 and 8, respectively. For the variable step size, the first step is equal to 4, comparable to experiment 1. Experiment 2 has been completed for starting point $(-3,-3,-3)$ and continuous noise only. For this starting point three tests were performed instead of a test and a retest only, used in experiment 1.

For experiment 1 the reliability of the multidirectional pattern search was determined. The reliability is the test-retest reproducibility for the pattern search. The reliability was calculated by the three-dimensional (3D) distance between the optima found in the test and in the retest condition. For experiment 2 two measures were determined. First the reliability was calculated for different pattern search designs that comprised different step size configurations. Second a convergence measure was determined. Convergence relates to the ability of the pattern search to find the optimum. The degree of convergence was studied by comparing the optima found for each step size configuration. That is, for each subject the average optimum setting (avg2) was determined for each step size configuration for starting point $(-3,-3,-3)$. After that the results were compared with the average optimum setting for the continuous noise condition and for all starting points of experiment 1 (avg1). Finally, the three-dimensional distance between avg1 and avg2 was calculated.

C. Subjects and test material

Experiment 1 was conducted by seven normal hearing subjects: s1 to s7. Only part of these subjects were available for experiment 2: s1, s3, and s4. Therefore, two other normal hearing subjects were added, s8 and s9, thus the data of experiment 2 rely on the results of five subjects. We verified that the performance of s8 and s9 was in agreement with the other subjects for a subset of conditions in experiment 1 (for details see the cluster analysis in the results section).

The subjects were listening with their ''better'' ear; three subjects used their left, six subjects their right ear. In the experiments, the subjects listened to speech in background noise. The duration of the fragments in the optimization procedure range between 1.5 and 2.5 seconds. We used standard Dutch sentences (VU98; see Versfeld *et al.*, 2000), uttered by a Dutch female speaker and equated for their root mean square (RMS)-values. Two background noises have been used: Continuous car noise (cont) and fluctuating speech-shaped noise (fluct, Festen and Plomp, 1990). The speech and noise were presented at a signal-to-noise ratio (SNR) of 0 dB. This SNR was assumed to be appropriate for judging the listening comfort of the speech-in-noise fragments used for the multidirectional search.[2]

D. Experimental setup

1. Recording method

The stimulus material was processed off-line. The speech fragments were preceded by a fixed period of 0.5 s of silence and the length of the silence period after the sentence depended on the length of the sentence, creating a fragment of four seconds. Likewise, four-second noise fragments were cut and added to the speech fragments at a signal-to-noise ratio of 0 dB. To achieve this, levels of the speech fragments were compared to the levels of the specific noise fragments in dBA. The A-weighting was used to correct for differences in spectral contents of the two noise types. After the speech and noise had been mixed, the signals were processed by the three algorithms.

The order of processing is quite straightforward. The noise reduction was used first, because spectral peaks are more pronounced for higher signal-to-noise ratios and may be processed more successfully after noise reduction. The spectral lift algorithm is applied last, after the enhancement stage, because the lift will be defined better once the signals are enhanced. When the order of processing is reversed, the amount of lift will depend on the amount of expansion. We carefully avoided clipping of the signals in the recording. To avoid loudness becoming a cue when fragments are compared pair wise in the experiments, the fragments were equalized with respect to their root-mean-square (RMS) values.

2. Playback method

Before running the optimization procedure the 44 100 Hz sampled signals were stored on disk. During the experiments these monophonic signals were transferred to a signal-processing hardware system (stage 1), filtered using an experimental hearing aid (stage 2), transferred back to the signal-processing hardware system (stage 3), and finally presented through headphones (stage 4).

In the first stage the monophonic signal was fed into channel 1 of the signal-processing hardware (Tucker Davis Technologies TDT system). An appropriate DC-voltage was fed into channel 2 of the signal-processing hardware system in order to create positive voltages needed for the experimental hearing aid in stage 2. After that, the signals of both channels were transferred via several hardware modules in different ways. That is, the AC-part, was converted to analog (using a TDT DA3-unit), filtered by an anti-aliasing filter (TDT FT5) with a cut-off frequency of 16 kHz and a slope of 48 dB/oct, and attenuated 18 dB by a programmable attenuator (TDT PA4). The DC-component was also converted to analog via the same converter (TDT DA3) and summed with the attenuated AC-component using a signal mixer (TDT SM3).

In stage 2, the signal with the appropriate DC-offset was transferred to a flexible experimental hearing aid (DASi-system; Rass and Steeger, 2000) to filter the signal following a predetermined frequency shape. We applied the correction factors described by Dillon (1986) to compensate for the presentation of signals by headphones instead of by a behind-the-ear hearing aid. In order to use this study as an appropriate reference for future studies with hearing-impaired subjects the normal hearing subjects in this study receive slight amplification for certain frequencies as well. The NAL-prescription rule was used for this purpose (National Acoustic Laboratory, Byrne and Dillon, 1986). In effect some frequency shaping was applied.

After this filtering, the signal was transferred back to the signal-processing hardware system in stage 3. In this stage the signal is amplified with a microphone amplifier (TDT MA2) and buffered by a headphone-buffer module (TDT HB6). Finally, in stage 4 the signal was presented monophonically by Telephonics TDH-39P headphones to the subjects seated in a sound-attenuating booth.

III. RESULTS

A. Pattern search optimization with adaptive step size

1. Statistics

The optimization procedure ended most often on meeting two stop criteria simultaneously. They were a setting is frequented three times and a setting has won a paired comparison more than six times. This might indicate that there is a close relation between these stop criteria. The total number of paired comparisons depended on subject and on noise type. To determine the significance of these differences, Wilcoxon (matched pairs signed rank sum) tests were performed. For continuous noise the number of paired comparisons was significantly lower (Wilcoxon, $p<0.05$) than for fluctuating speech-shaped noise, on average they are 30 and 38, respectively. In the paired comparisons the second alternative is used not significantly more frequently than the first alternative, although the second alternative tends to be used more often.

The relation between use of the ''repeat'' and ''no difference'' button was also subject dependent. Most subjects did not use the ''no difference'' button frequently. So, these subjects did not seem to have frequent problems in distinguishing the fragments, irrespective the step size. Two subjects used the no difference button relatively often, probably

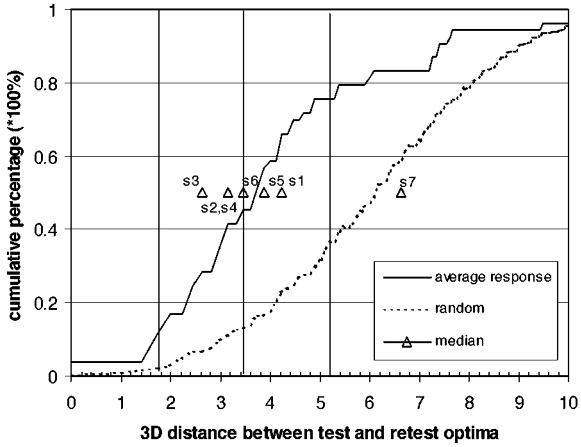

FIG. 1. Cumulative distributions for the calculated 3D distances between the test and retest optima, averaged for two noise types and seven normal hearing subjects (average response). As a reference, the random distribution is also given as calculated based on randomly assigned test and retest optima. The triangles represent the median 3D distance values for each individual (s1 to s7).

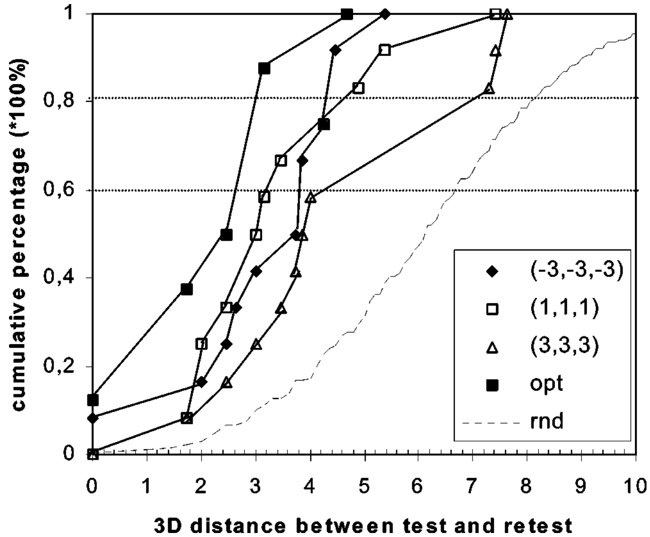

FIG. 2. Cumulative distributions for the calculated 3D distances between the test and retest optima, averaged for six subjects and differentiated by initial setting. The initial setting was (−3,−3,−3), (1,1,1), (3,3,3), or the average optimal setting (opt). As a reference, the random distribution is also given as calculated based on randomly assigned test and retest optima.

because they sometimes had difficulties hearing differences between the fragments. One subject almost always used the repeat button when there was no perceived difference. For the fluctuating noise relative to the continuous noise, the repeat button was used twice as often and the no difference button a factor four more often.

After a startup round with a step size of 4, the step size was halved to 2. From this point on the step size was recorded to determine the minimum step size and maximum step size used in the search. The minimum step size could either be 1 or 2. The maximum step size could be either 2 or 3. The results showed that the minimum step size per dimension was comparable for both noise types and had a value of 1 for approximately 80% of all paired comparisons. However, the maximum step size was noise dependent. A maximum step size of 3 was found in 7% and 27% of the cases, for continuous noise and fluctuating noise, respectively. This large difference was found to be mainly caused in the spectral enhancement dimension. Obviously, spectral enhancement settings may be harder to discriminate in fluctuating noise.

2. Reliability

To acquire information about the reliability we used two strategies. Cumulative distributions were used to determine the factors that influence the reliability. A cluster analysis provided information of the spread and location of optimal and suboptimal settings for each individual.

a. Cumulative distributions. The three-dimensional (3D) distance between the optimum found for the test and retest condition has been calculated in order to determine the reliability of the optimization procedure. In principle, the procedure has maximum reliability when the test and retest results always yield the same optimum settings for all conditions. Cumulative distributions were used in order to analyze the spread of possible test-retest differences. The cumulative percentages were calculated as a function of the maximum distance between test and retest optima. These percentages are given in Fig. 1. From this graph it appears that 11%, 45%, and 76% of the data have a test-retest distance smaller than $\sqrt{3}$, $2\sqrt{3}$, and $3\sqrt{3}$, respectively. These distances represent a maximum step size of 1, 2, or 3, in each dimension. As a reference the random distribution is also given, presented as a dashed line. For this random distribution 1000 test and retest results have been chosen randomly, after which the 3D distances have been calculated. The cumulative distribution of the normal hearing subjects deviates clearly from the random cumulative distribution. The cumulative distribution of the normal hearing subjects contains a few distances that are very large. This is mainly due to the results of subject 7. To illustrate the large test–retest differences, we have plotted the median 3D distance for each individual (see Fig. 1), which shows that the performance of s7 was very close to completely random.

To reliably study the dependence of the reliability on noise type and initial setting the results of s7 have been discarded. For different noise types, the cumulative distributions were found to be approximately equal. For different initial settings, however, differences in the distributions appeared, as shown in Fig. 2. Especially for percentages around 80% differences occur. In general, the test-retest reliability was highest when the average optimal setting (opt) was chosen as the initial setting of the procedure. Initial setting (3,3,3) produced the lowest reliability.

The one-dimensional (1D) distances were calculated for each separate dimension. These results are presented in Fig. 3. It appears that the test-retest reliability depends strongly on the algorithm. The reliability was highest for the noise reduction algorithm and poorest for spectral lift. Additionally, the one-dimensional reliability depended on noise type. The reliability for SE and NR was highest for continuous noise, whereas for SL the reliability was highest for fluctuating speech-shaped noise (not shown in Fig. 3).

b. Cluster analysis. A cluster analysis has been performed to investigate the spread of optima found for differ-

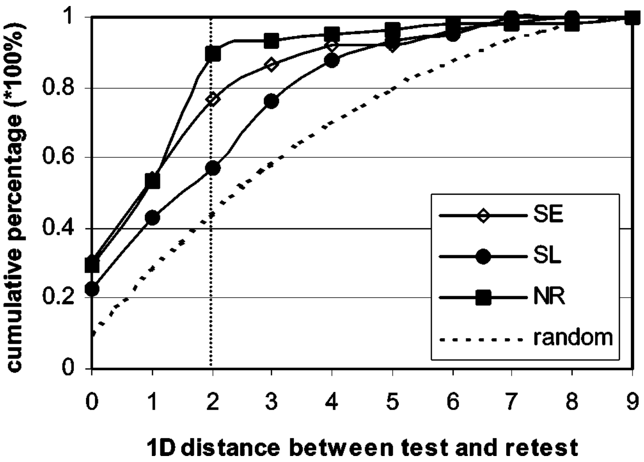

FIG. 3. Cumulative distributions for the calculated 1D distances, i.e., the difference between test and retest for each dimension separately. The dimensions of the optimization procedure represent the algorithms Spectral Enhancement (SE), Spectral Lift (SL), and Noise Reduction (NR). For comparison, the random distribution is also given.

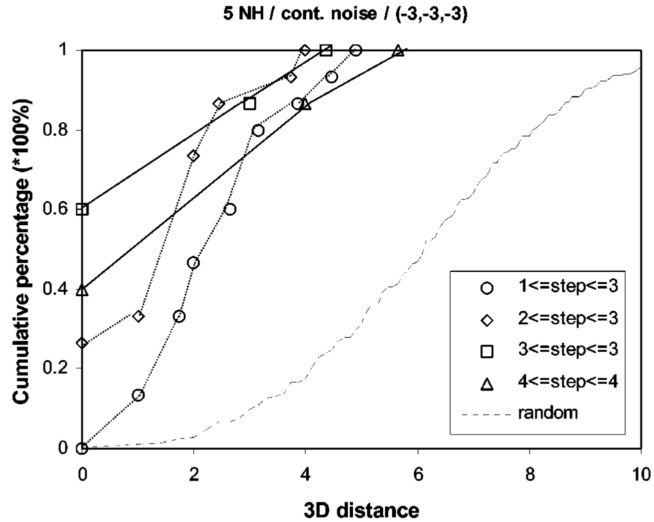

FIG. 4. Cumulative distributions for the calculated 3D distances between the test and retest optima, averaged for five normal hearing subjects (5NH) and differentiated by step-size configuration. The pattern searches were performed for continuous noise (cont. noise) and initial setting (−3,−3,−3) only. The step size can be either adaptive and vary between 1 and 3 ($1 \leqslant step \leqslant 3$), or between 2 and 3 ($2 \leqslant step \leqslant 3$), or the step size can be fixed at 3 ($3 \leqslant step \leqslant 3$) or fixed at 4 ($4 \leqslant step \leqslant 4$). As a reference, the random distribution is also given as calculated based on randomly assigned test and retest optima.

ent conditions in the three-dimensional space of algorithm settings. The method of hierarchical clustering has been used, based on nearest neighbors and Euclidean distances. The results of the cluster analysis are presented in Table I. The numbers of final settings that are included in the cluster analysis are indicated in the second column (total). The criterion to assign different optima to the same cluster is the average distance between the cluster center and the other points, which should be lower than $\sqrt{3}$. The number of clusters that result from the analysis with this criterion is given in the third column (# clusters). The largest cluster is assumed to be the optimum. The one but largest cluster is considered a suboptimum. The algorithm settings that represent the optimum or suboptimum are rounded values of the cluster center. The number of points (# points) that form the optimum and suboptimum are also included in Table I. Clusters that contain only one or two final settings are considered to be outliers and are given in the last column (# outliers).

For subjects s1 to s6 at least 50% of the final settings is grouped in one cluster. So, for the majority of the subjects the optimization strategy can find an optimal combination of algorithm settings. For s1, s4, s5, and s7 there is also a second cluster of reasonable size. For s1 and s4 this is a true suboptimum, i.e., a serious competitor for being most comfortable. The presence of two large clusters for subjects s5 and s7 appear to be optima that relate specifically to the two noise conditions.[3] The percentage of outliers is 25% or less for all but one subject. Subject s7 has a large scatter among the final settings found by the optimization procedure.

There proved to be a clear relation between the number of paired-comparisons needed and the size of the clusters. For the large and medium optimum and suboptimal clusters the procedure stopped on average after 25 and 30 paired comparisons, respectively. For the outliers, 54% of the procedures that were performed took 45 to 54 paired comparisons, which is close to, or at, the upper limit of paired comparisons needed.

TABLE I. Results of a cluster analysis for seven subjects (s1 to s7). The total number of points (total) included in the analysis and the number of clusters (#clusters) that result from the analysis are given in the second and third column. The two largest clusters are assumed to form the optimum and suboptimum. The algorithm settings of the optimum and suboptimum are indicated. SE, SL, and NR represent algorithms Spectral Enhancement, Spectral Lift, and Noise Reduction. Clusters that contain only one or two final settings are called outliers. The number of points (#points) are given for the optimum cluster, the suboptimum cluster and the clusters of outliers.

			Cluster1: optimum				Cluster2: suboptimum				Outliers
Subject	Total	#Clusters	SE	SL	NR	# points	SE	SL	NR	# points	# points
s1	16	4	−1	5	0	11	0	1	0	3	2
s2	16	4	1	4	0	12					4
s3	16	3	−1	−1	0	14					2
s4	16	6	0	1	0	8	0	−3	1	4	4
s5	14	4	0	3	2	7	−1	−2	1	4	3
s6	14	4	−1	1	1	10					4
s7	16	8	5	−3	5	5	−1	−3	6	4	7

B. Pattern search with step size as parameter

The results of all four step-size configurations (step123, step23, step3, step4) are plotted as cumulative distributions in Fig. 4. For comparison, the random distribution, based on randomly assigned test and retest optima, is plotted in the dashed line.

The results show that the reliability of the procedure tends to increase when the minimum step size increases from one to three (for step123, step23, and step3, respectively). For a fixed step size of four the reliability decreases relative to a step size of 3. Both for the fixed (solid lines) and for the variable step size configurations (dotted lines) there is a parallel shift of the distributions. This shift indicates an increase in reliability for the variable configurations and a decrease for the fixed configurations when the step size is enlarged.

The degree of convergence was studied by comparing the optima found for each step size configuration. The average results indicate that the 3D distance increases with minimum step size, at least for minimum step sizes between 1 and 3. The 3D-distance is 1.0, 1.8, 2.0, and 1.7 for configurations step123, step23, step3, and step4, respectively. There is a significant difference ($p<0.05$, Wilcoxon matched pairs signed rank sum test for the results of five subjects and four configurations) between step123 and the step3 configuration, which indicates that the optimization procedure converges better for the variable step size configuration.

Finally, the number of paired comparisons needed to end the optimization procedure appeared inversely proportional to the minimum step. For step123, step23, step3, and step4 the number of paired comparisons is 22.2, 16.2, 16.2, and 13.8, respectively.

IV. DISCUSSION

A. Adaptive pattern search

An advantage of the pattern search method is that no a priori information is required. For perceptually governed optimization that depends on the listener's capacities and experience this method shows promising capabilities. However, the perceptual capacities of the listeners also constrain the methodological possibilities of the optimization method. The adaptive step size that is used in this study helps to create perceptually distinguishable fragments to be used in the paired comparisons. The results of the pattern search procedure suggest that settings that differ 2 or 3 steps can be judged reliably for our set of stimuli. That is, the reliability of the optimization procedure is high for a variable step size that ranges between 2 and 3, and for a fixed step size of 3.

The optimal settings have been estimated in two different ways considering all pattern searches for each listener. The estimators comprise the average optimum setting (O_{avg}) and the center of the largest cluster (Cl). The three-dimensional distance between the two estimators can be determined. The individual results, average results, and standard deviations are presented in Table II for each background noise type. The results of this study suggest that for both noise types the correspondence between the cluster center and the average optimum setting is close for all but one subject. This can be seen from the small average 3D-distances that are smaller than $\sqrt{3}$ for all but one subject (distances larger than $\sqrt{3}$ are presented in gray in Table II). Although the results presented in Table II cannot simply yield information about the quality of the estimators, the cluster center of the largest cluster is assumed to be the best estimator of the optimum. For in some pattern searches (for different tests and starting points) the ''real'' optimum can be missed when there are grave discontinuities in the three-dimensional perceptual landscape or due to procedural inaccurateness as a result of step sizes used.

TABLE II. Three-dimensional distances between the cluster optimum (Cl) and the average optimum (Oavg) differentiated by subject (s1 to s7) and noise type. The noise type was either continuous car noise (cont) or fluctuating speech-shaped noise (fluct). Additionally, the average three-dimensional distances (avg) and standard deviations (stdev) are indicated. Values larger than $\sqrt{3}$ are presented in italic.

Subject	3D-dist(Cl, Oavg) cont	3D-dist(Cl, Oavg) fluct
s1	1.3	*3.1*
s2	0.2	0.3
s3	0.2	0.3
s4	*3.2*	0.6
s6	0.3	0.5
s7	0.4	0.5
avg	0.9	0.9
stdev	1.2	1.1

B. Reliability and convergence

The cumulative distributions of all subjects except for s7 indicate that the three-dimensional distance between test and retest is lower than $2\sqrt{3}$ for about 50% of the results. The cluster analysis illustrates the presence of large and intermediate clusters around the optimum, small clusters around a suboptimum, and some outliers.

In studies of Neuman *et al.* (1987), Kuk and Pape (1992), and Kuk and Lau (1996) the reliability of the modified ''Simplex'' procedure was higher than in this study. The most important explanation for that may well be connected to the configuration and the stimulus material of the optimization procedure. In our setup the subjects listened to sentences that were between 1.5 and 2.5 s in duration. Moreover, in our experimental paradigm speech-in-noise fragments were presented in succession and could be repeated only once. A problem of this design is that the auditory memory is not sufficient to compare the fragments: At the end of fragment 2, a part of fragment 1 is already forgotten. In the studies mentioned above the paradigm is very different in that continuous discourse is used and listeners are allowed to switch freely between two parameter settings. By switching freely, the effects of auditory memory are less acute, because of the possibility to return to the alternative. In a small parameter or algorithm space such a paradigm appears to be convenient, but for large multidimensional setups, like in this study, such a slow approach is unpractical.

Another, maybe more influential difference is the level equalization of the stimulus material in our study. For the study of Neuman *et al.* (1987) the loudness of low frequencies and high frequencies of the stimulus material was one of the optimization parameters under study. For the studies of Kuk and Pape (1992) and Kuk and Lau (1996), the parameter settings of the two-dimensional optimization procedure were the degree of amplification used for high and low frequencies. However, because the subjects had to adjust the level of the test material to their most comfortable level (MCL), the effective levels can be quite different for different settings. The advantage of these approaches is that the reliability is high because level differences are relevant cues for converging on the optimum setting. An additional advantage is that (parts of) the stimulus material can never be too loud or too soft. Our optimization procedure was intended to focus on the effects of the algorithms themselves and not on level differences. The reason for this is that loudness differences can act as a separate and different evaluation criterion. The procedure used in this study avoids that the optimum setting found is the result of a loudness judgment exclusively.

The choice of the background noise type in combination with the selected signal processing algorithms also affects the reliability. An explanation could be that the differences between settings, in the paired comparisons in the optimization process, are much more difficult to detect for the fluctuating noise than for the continuous noise. Subjects indicated that processing artifacts are much more prominent for the continuous car noise condition relative to the fluctuating speech-shaped noise condition.

When the maximum number of paired comparisons was reached the end point is not likely to be the optimum, and the setting that has won most paired comparisons is in most cases not the optimal setting either. Conversely, when the end point is reached rather quickly, within nine to eighteen paired comparisons, this point is much more likely to be the true optimum. The reliability found for each noise type can be directly related to the stop criterion, if addressed. For fluctuating noise, that had a relatively low reliability, the procedure stopped most often when the maximum number of paired comparisons was reached. For continuous noise, the procedure ended both on three visits and more than six wins. This might imply that combining these both stop criteria yields highest reliability.

The adaptive pattern search of study A (step-size configuration step123) appeared to have the best convergence properties. The unexpected reversal with respect to the degree of convergence as a function of step size might be due to the case that a grid point was accidentally located near avg1. Further inspection of the data revealed that this was the case in only one subject. In three out of five subjects avg1 proved to be closer to a grid point in the step3 configuration than in the step4 configuration. So, this effect is not very likely to be the only explanation.

C. Auditory constraints

In this study we did not record the difference limens for each particular dimension. Although these limens can be measured for each algorithm separately, that would still not produce any knowledge of the limens for combinations of these nonlinearly acting algorithms. The present study shows that the adaptive step size can partly compensate for any unfortunate choices of the step size. The step size configuration appears to affect the reliability. Apart from the fixed step size configurations, we found that the highest reliability was achieved for a minimum step size of two and a maximum step size of three (step23). A variable step size helps to create perceptually distinguishable fragments. In the same line, the elevated minimum step size has an advantageous effect in avoiding the presentation of indistinguishable fragments.

In the studies of Neuman *et al.* (1987) and Kuk and Pape (1992), it seems that the possible existence of multiple optimum settings played a minor role. That is, the combination of the parameters under study yielded a response surface that curved smoothly around a single optimum. Yet, in the study of Kuk and Lau (1996) category rating results showed a multimodal response pattern for three out of seven subjects, when evaluated for clarity for SNR=0 and SNR=5 dB. Also the results of a previous pilot study indicate that the response surface can sometimes be multimodal. In this study, the cluster analyses imply the presence of multiple optima.

V. CONCLUSIONS

Inclusion of an adaptive step size has two advantages. The perceptibility between fragments can be controlled. That is, an increase in the step size can be beneficial in order to create perceptual differences between settings that are compared. As a result the steps in the procedure can be estimated more reliably. The gradual decrease of the step size helps the procedure to converge towards the optimum setting. The reliability, however, depends on many factors, especially on initial value, stop criterion, step size constraints, background noise, algorithms used, and the possible presence of suboptimal settings. Reliability and convergence appear to have a trade-off effect. Specifically, when the step size is enlarged the reliability improves, but the convergence deteriorates.

The results of the cluster analysis and of the cumulative distributions correspond closely, especially for continuous car noise, with the exception of one subject. The cumulative distributions indicate that for the pattern search about 50% of the test-retest optima have a three-dimensional distance lower than $2\sqrt{3}$. The cluster analyses show that large clusters can be formed that cover many end points of the pattern search found for different conditions. For four subjects a second cluster could be identified that either represented the optimum for the second background noise, or a suboptimum.

The presence of a suboptimum can cause the optimization strategy to stop prematurely. Another drawback of the present pattern search procedure is the problematic division between "not distinguishable" and "no preference." Nevertheless, the simplicity and efficiency of the search and the assimilation of algorithm interactions make the optimization strategy suitable for fine-tuning of auditory signal-processing algorithms.

ACKNOWLEDGMENTS

This study was supported by the Heinsius-Houbold Fund. We like to thank László Körössy of the Academic Medical Center Amsterdam for his technical support.

[1]Using the paired-comparison paradigm the expression "Simplex method" might be confusing. A "Simplex" designates a Euclidean geometrical spatial element bounded by a minimum number of points. For two dimensions a Simplex is a triangle, in three dimensions a tetrahedron. So for n dimensions a Simplex contains $n+1$ vertices (boundary points). As such, the expression Simplex can be used, because the settings that are compared pair wise correspond to vertices of a Simplex configuration. In the procedure new Simplices are formed by reflecting one point in the center of gravity (centroid) of the remaining points. This so-called Simplicial feature of the Simplex method is, among others, fundamentally different from the (modified) Simplex procedure of Neuman *et al.* (1987). When using that procedure in a paired-comparison paradigm, it is impossible to decide what vertex should be reflected. This meant that different procedural rules had to be incorporated. Instead of reflection in one point, as is the case in the original procedure, the new Simplex is created by a combination of translations and line-reflections. As a consequence, the new Simplex can share at most 1 point with the old Simplex, while in the procedure of Spendley *et al.* (1962) the new Simplex always shares $n-1$ points with the old Simplex. This procedural difference indicates that the original elegant movement of the Simplex was abandoned in the modified procedure of Neuman. To avoid ambiguous terminology, the most important features of our (and Neuman's) procedure have been selected and compared with other procedures that can be classified as direct search methods (see e.g., Torczon, 1989; Lewis *et al.*, 1998, 2000). To appreciate the lattice structure and the constrained search directions, it is convenient to denote our procedure "multidirectional pattern search."

[2]In order to check this assumption, the Speech Reception Thresholds (SRT) of six subjects were measured. This threshold represents the signal-to-noise ratio at which 50% of sentences in noise are repeated entirely correctly (Plomp and Mimpen, 1979). On average, the subjects' SRT is -3.6 (1.6) and -9.1 (2.5) dB for two standard noises, continuous speech-shaped noise and fluctuating speech-shaped noise, respectively. This corresponds well with results of Festen and Plomp (1990) and is substantially lower than the SNR of 0 dB for the speech-in noise fragments used in this study. The SNR for continuous car noise, which was not measured here, can be expected to be lower than the SNR for continuous speech-shaped noise due to the fact that continuous speech-shaped noise will mask speech more efficiently than continuous car noise. So, the paired comparisons were performed for speech in noise presented at levels for which the percentage correct score was well over 50%.

[3]When for each noise type cluster analyses were carried out for each individual, the same conclusion could be drawn. Moreover, it appeared that the number of optima that form the largest cluster and the number of outliers were comparable for subjects s8 and s9 and the other seven subjects. Hence, it is assumed that the results of the five subjects in experiment 2 can be compared safely amongst each other and with the results of experiment 1.

Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**(4), 257–265.

Dillon, H. (**1997**). "Converting insertion gain to and from headphone coupler responses," Ear Hear. **18**(4), 346–348.

Festen, J. M., and Plomp, R. (**1990**). "Effect of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**(4), 1725–1736.

Kuk, F. K., and Pape, N. C. M. (**1992**). "The reliability of a modified simplex procedure in hearing aid frequency-response selection," J. Speech Hear. Res. **35**, 418–429.

Kuk, F. K., and Lau, C. (**1996**). "Comparison of preferred frequency gain settings obtained with category rating and modified Simplex procedure," J. Am. Acad. Audiol **7**, 322–331.

Lewis, R. M., Torczon, V., and Trosset, M. W. (**1998**). "Why pattern search works," Optima (Mathematical Programming Society Newsletter), 1–7.

Lewis, R. M., Torczon, V., and Trosset, W. (**2000**). "Direct search methods: then and now," J. Comput. Appl. Math. **124**, 191–207.

Lyzenga, J., Festen, J. M., and Houtgast, T. (**2002**). "Speech enhancement scheme incorporating spectral expansion evaluated with simulated loss of frequency selectivity," J. Acoust. Soc. Am. **112**(3), 1145–1157.

Neuman, A. C., Levitt, H., Mills, R., and Schwander, T. (**1987**). "An evaluation of three adaptive hearing aid selection strategies," J. Acoust. Soc. Am. **82**(6), 1967–1976.

Plomp, P., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.

Rass, U., and Steeger, G. H. (**2000**). "A high performance Pocket-size System for Evaluations in Acoustic Signal Processing," Acta Acust. (Beijing) **86**, 374–375.

Spendley, W., Hext, G. R., and Himsworth, F. R. (**1962**). "Sequential Application of Simplex Designs in Optimisation and Evolutionary operation," Technometrics **4**, 441–461.

Torczon, V. J. (**1989**). "Multi-directional search: a direct search algorithm for parallel machines," Thesis for philosophy doctorate, Rice University, Houston, Texas.

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**(3), 1671–1684.