Using Genetic Algorithms with Subjective Input from Human Subjects: Implications for Fitting Hearing Aids and Cochlear Implants

Deniz Başkent, Cheryl L. Eiler, and Brent Edwards

Objective: **To present a comprehensive analysis of the feasibility of genetic algorithms (GA) for finding the best fit of hearing aids or cochlear implants for individual users in clinical or research settings, where the algorithm is solely driven by subjective human input.**

Design: **Due to varying pathology, the best settings of an auditory device differ for each user. It is also likely that listening preferences vary at the same time. The settings of a device customized for a particular user can only be evaluated by the user. When optimization algorithms are used for fitting purposes, this situation poses a difficulty for a systematic and quantitative evaluation of the suitability of the fitting parameters produced by the algorithm. In the present study, an artificial listening environment was generated by distorting speech using a noiseband vocoder. The settings produced by the GA for this listening problem could objectively be evaluated by measuring speech recognition and comparing the performance to the best vocoder condition where speech was least distorted. Nine normal-hearing subjects participated in the study. The parameters to be optimized were the number of vocoder channels, the shift between the input frequency range and the synthesis frequency range, and the compression-expansion of the input frequency range over the synthesis frequency range. The subjects listened to pairs of sentences processed with the vocoder, and entered a preference for the sentence with better intelligibility. The GA modified the solutions iteratively according to the subject preferences. The program converged when the user ranked the same set of parameters as the best in three consecutive steps. The results produced by the GA were analyzed for quality by measuring speech intelligibility, for test-retest reliability by running the GA three times with each subject, and for convergence properties.**

Results: **Speech recognition scores averaged across subjects were similar for the best vocoder solution and for the solutions produced by the GA. The average number of iterations was 8 and the average convergence time was 25.5 minutes. The settings produced by different GA runs for the same subject were slightly different; however, speech recognition scores measured with these settings were similar. Individual data from subjects showed that in each run, a small number of GA solutions produced poorer speech intelligibility than for the best setting. This was probably a result of the combination of the inherent randomness of the GA, the convergence criterion used in the present study, and possible errors that the users might have made during the paired comparisons. On the other hand, the effect of these errors was probably small compared to the other two factors, as a comparison between subjective preferences and objective measures showed that for many subjects the two were in good agreement.**

Conclusions: **The results showed that the GA was able to produce good solutions by using listener preferences in a relatively short time. For practical applications, the program can be made more robust by running the GA twice or by not using an automatic stopping criterion, and it can be made faster by optimizing the number of the paired comparisons completed in each iteration.**

(Ear & Hearing 2007;28;370–380)

Many modern hearing aids and cochlear implants offer numerous features, in addition to providing basic audibility, that have to be optimized for an individual user. Finding the optimal settings can be difficult, as individuals might have different pathologies in the auditory system and might also have different listening preferences (Preminger & Van Tasell, 1995). Moreover, some of the features might interact with each other, further complicating the fitting process. Theoretically, the best settings can be determined by functional measurements that can be made for each patient and for all device features, individually or in combinations. However, this would not be realistic as such a fitting would require more time and expense than most clinics or patients could afford. To simplify the fitting process for clinicians, manufacturers provide default parameter settings based on clinical and electro-acoustic data, and the best parameter values for each listener are usually found by trial-and-error. This limited set of parameters might not be sufficient to provide a satisfactory fitting to all patients with varying pathologies and preferences. Furthermore, with the advances in digital signal processing and features that are becoming more sophisticated, manufacturers themselves

Starkey Hearing Research Center, Berkeley, California.

0196/0202/07/2803-0370/0 • Ear & Hearing • Copyright © 2007 by Lippincott Williams & Wilkins • Printed in the U.S.A.

might not be fully aware of the best default settings for new algorithms.

For fitting the gain, the most fundamental feature of a hearing aid, prescriptive formulas have been developed. Conventional gain prescriptions, such as NAL-R (National Acoustic Laboratories-Revised; Byrne & Dillon, 1986), NAL-NL1 (NAL nonlinear; Byrne et al., 2001), POGO (Prescription of Gain and Output; McCandless & Lyregaard, 1983), and Berger's method (Berger et al., 1977), are based on measurements of audiometric thresholds only. These methods are fast; however, they do not accommodate preferences of individual users and may need further tuning to improve patient satisfaction (Byrne & Cotton, 1988; Kuk & Pape, 1992). Some other prescriptive methods, such as LGOB (Loudness Growth in ½-Octave Bands; Allen et al., 1990) or IHAFF (International Hearing Aid Fitting Forum; Valente & Van Vliet, 1997), include a loudness measure in the fitting to customize the loudness growth for individual users. There are also adaptive gain fitting procedures, such as Scaladapt (Kiessling et al., 1996), Camadapt and Ear Tuner (Moore et al., 2005), that are used for finding optimal settings of multichannel compression hearing aids for individual users. In contrast to prescriptive methods, the adaptive procedures provide a highly functional fitting for individuals by systematically adjusting the hearing aid settings according to a patient's preferences for loudness, sound quality and/or comfort, under a range of listening conditions.

The adaptive procedures have been shown to be beneficial for individual users; however, they were specifically developed for fitting compression only. Other customization methods were suggested that could be used for optimization of different device features. One suggestion was to give more control to the patient. McDermott (1998) and Zakis (2003), for example, designed portable sound processors that can be connected to hearing aids. With these, the patient can register the preferred settings for different listening conditions. Such a processor can be used to dynamically change the device settings or to keep track of patient preferences that can later be used in the clinic for a more efficient fitting. However, if no guidance is provided, it might be an overwhelming task for the patient to assess each of the numerous settings that the hearing aid offers.

Optimization algorithms have been proposed for a fast, systematic, and flexible fitting of device parameters. A modified simplex algorithm was used for fitting gain in hearing aids (Kuk & Pape, 1992; Neuman et al., 1987; Preminger et al., 2000; Stelmachowicz et al., 1994). Genetic algorithms (GA) were used for fitting features related to hearing aids (Durant et al., 2004) and cochlear implants (Bourgeois-République et al., 2005; Wakefield et al., 2005). These algorithms produce candidate parameter settings that are evaluated by the patient who listens to speech stimuli with the device under each setting. A set of the device parameters is modified according to the rules of the optimization algorithm using the subjective input of the patient. These steps of evaluation and modification continue in iterations until parameter settings that are satisfactory to the patient are found. Optimization algorithms are generally fast because the final solution is usually reached by evaluating only a small fraction of all possible solutions. Flexibility is another advantage as they do not have to be limited to gain fitting only; any device feature can be customized with these algorithms. Franck et al. (2004), for example, used the modified simplex algorithm to optimize parameters related to noise reduction, spectral enhancement, and spectral tilt, while all working in conjunction. Durant et al. (2004) used the GA for fitting parameters of feedback cancellation. Wakefield et al. (2005) used the GA for finding the optimal settings for many cochlear implant parameters simultaneously, including the number of active electrodes and stimulation rate of electrical pulses sent to the electrodes, the number of maxima used in the speech-processing strategy.

Difficulties exist with applications involving input from human subjects (Takayagi, 2001). In conventional applications of optimization algorithms, there is usually an output metric for the system to be optimized. In a communication system, for example, this can be the distortion in the transmitted signal. In this case, the solutions offered by an optimization algorithm can be evaluated by direct measurement of the distortion. When the optimization algorithms are used for fitting settings to a human listener's preferences, however, the main evaluation tool is the subjective response from the listener. Usually, there is no metric available to quantitatively measure the suitability of the final solution. This is, in fact, one of the reasons why simplex or genetic algorithms were suggested for perceptual optimization; these algorithms do not require an analytical expression that describes the possible solutions, which is required by many other optimization algorithms such as the conjugate gradient descent.

It is crucial that the feasibility of an optimization program is assessed objectively before it can be suggested for real life applications, where an objective assessment of the final solution might not be available. In the present study, the feasibility of GAs in optimizing auditory settings using the subjective input from listeners is analyzed comprehensively. To produce a listening problem with an output metric, speech was intentionally distorted using a noiseband vocoder (Shannon et al., 1995). Three

parameters of the vocoder processing were used for optimization; number of spectral channels, shift between the input frequency range and the synthesis frequency range, and compression-expansion of the input frequency range over the synthesis range. These parameters were selected because the acute effects of these spectral manipulations on intelligibility of speech by normal-hearing subjects were known from previous studies (Başkent, 2006; Başkent & Shannon, 2003, 2007; Friesen et al., 2001; Fu & Shannon, 1999), so the final solutions produced by the GA could be quantitatively evaluated by using similar speech recognition tests. The subjects listened to distorted speech and entered their preferences in the GA in paired comparisons. We hypothesized that if subjects could make a reliable judgment of speech intelligibility, as was shown to be the case, in general, by Punch & Parker (1981), and if the GA was an appropriate tool for optimizing parameters for the best listening conditions, speech recognition with the optimal settings produced by the GA should be high. In addition to performance with the GA solutions, the data were also analyzed for convergence and repeatability, as these factors would also be of importance for practical applications.

METHODS

Subjects

Nine normal-hearing listeners between the ages of 19 to 34, with an average of 24.3 yr, participated in the experiment. All subjects were native speakers of American English and had air conduction thresholds better than 20 dB HL at audiometric frequencies ranging from 250 to 6000 Hz bilaterally. The immittance test results from tympanograms and acoustic reflex thresholds were consistent with normal middle ear function in both ears.

Stimuli

For the practice session, TIMIT sentences (Garofolo et al., 1993) were used. These sentences are not phonetically balanced and are relatively difficult compared to other sentence databases. The sentences were spoken by multiple talkers with different dialects. For the GA runs and speech recognition tests, IEEE sentences (IEEE, 1969) were used. The IEEE database includes 720 sentences of similar length and phonemic content. The sentences were spoken by a male speaker.*

Noiseband Vocoder Processing

A method widely used to systematically explore the effects of temporal and spectral degradations on speech perception is the noiseband vocoder (Dudley, 1939; Shannon et al. 1995; Xu et al., 2005). Processed speech at the output of the vocoder is a sum of narrow bands of noise (carrier bands) that were modulated with envelopes extracted from individual bands of speech (analysis bands). As a result, only the crude spectral and temporal elements of the input speech are retained. The noiseband vocoder has also been used to simulate cochlear implant processing with normal-hearing listeners (Green et al., 2005; Poissant et al., 2006). In the simulations, the carrier noise bands of the vocoder represent the stimulation range in the cochlea, determined by simulated electrode locations, while the analysis bands represent the acoustic input.

In the present study, the analysis bands were produced by filtering sentences into frequency bands using 6th-order Butterworth bandpass filters. The filter cutoff frequencies were determined by (1) converting the input frequency range in Hertz into cochlear distance in millimeters using Greenwood's cochlear mapping function (Greenwood, 1990), (2) dividing the entire range in mm into equal cochlear distances, and (3) converting the distances back into the frequency domain using the mapping function. The speech envelope was extracted from each analysis band by half-wave rectification and low-pass filtering, using a 3rd-order Butterworth filter with a cut-off frequency of 160 Hz (at –3 dB). The noise carrier bands were obtained by filtering wideband noise with a second set of 6th-order Butterworth filters. In the final stage, the noise bands were modulated with the envelopes, and all modulated noise bands were summed to produce the processed speech. The amplitude levels were adjusted such that the original and processed tokens had the same overall RMS energy. No additional manipulation was performed to match the spectral shape.

In the present study, a carrier band range of 16 mm in cochlear distance was selected for consistency with previous studies (Başkent & Shannon, 2003, 2007; Fu & Shannon, 1999). The first vocoder parameter used in the GA, number of channels, was varied by changing the number of bandpass filters used, as shown in the left column of Figure 1. The second parameter, the spectral shift, was implemented by shifting the carrier band range to lower or higher frequencies, as shown in the middle column of Figure 1. The third parameter, compression-expansion, was implemented by making the analysis band range wider or narrower than the carrier band range, as shown in the right column of Figure 1. The setting with maximum number of channels and least spectral distortions, shown in the top left corner of Figure 1, represented the optimal solution. Lowering the number of channels or introducing

*The recordings were made by Dr. Qian-Jie Fu and John Galvin at the House Ear Institute, Los Angeles, CA.

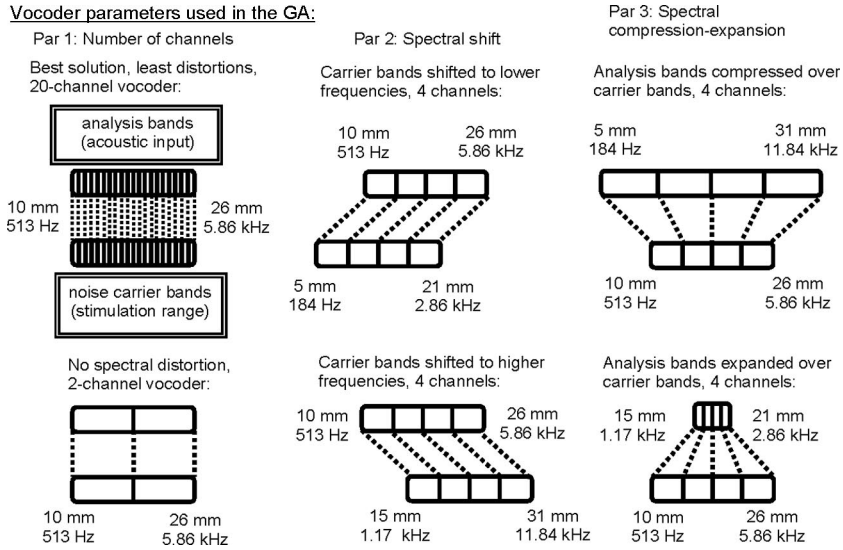

Fig. 1. Spectral manipulations shown for the noiseband vocoder parameters of number of channels, spectral shift, and spectral compression-expansion. For each manipulation, upper bands show analysis bands and lower bands show noise carrier bands. First column shows the effect of varying number of channels, with no shift or compression. Second column shows the shift manipulation where the carrier bands were shifted to lower or higher frequencies. Third column shows the manipulation where the analysis bands were wider or narrower than the carrier bands, resulting in spectral compression or expansion, respectively.

spectral distortions were expected to cause poorer speech intelligibility compared with the optimal solution. The ranges of values used for each parameter are summarized in Table 1. With these values, the parameter set with the highest speech recognition performance was expected to be the least distorted vocoder condition of [20 0 0].

Genetic Algorithm

The GA is an optimization method based on concepts borrowed from evolution, such as the survival of the fittest, and the stochastic operators of mutation and cross-over (Mitchell, 1997). One set of parameters that is being optimized by the GA is called a gene. In contrast to many optimization algorithms, the GA works on a population of genes, rather than an individual set of parameters. In each iteration, the population of genes is evaluated for fitness, and the genes are modified accordingly to produce the next generation of genes using one of these methods: (1) Elitism: A gene, usually with high fitness value, is passed onto the next generation with no alterations; (2) Mutation: Some or all parameters of a gene, usually randomly selected from the population, are changed by random values; (3) Cross-over: Two parent genes, randomly selected from the population, produce new child genes by exchanging individual parameter values or averaging randomly selected parameters from each parent. In the next iteration, the new fitness values are determined for the new generation of the genes. The iterations are repeated until a convergence criterion is satisfied.

TABLE 1. Lower and upper limits and the step sizes of the vocoder parameters used in the genetic algorithm

Vocoder parameters	Lower limit	Upper limit	Step size
Vocoder channels	2 channels	20 channels	1 channel
Frequency shift	−8 mm	+8 mm	1 mm
Frequency compression-expansion	−6 mm	+8 mm	1 mm

In the present study, every gene was a combination of three vocoder parameters: number of vocoder channels, the spectral shift between the analysis and carrier band ranges (expressed as cochlear distance), and the spectral compression-expansion of analysis filter range over the carrier filter range (expressed as cochlear distance). The population consisted of six genes, a value selected to be optimal for the present study. For most GA applications, it is beneficial to have a large number of genes, as the ability of the GA to find the optimal solution is also related to the number of the genes. However, in the present study, a large population size would slow down the overall program considerably, as the subjects would need more time to evaluate all genes.

At the beginning of the program, the genes were produced randomly, within the limits of each parameter, as shown in Table 1. With the selected step sizes and the lower and upper limits of the parameters, there were 4845 possible discrete settings.

The fitness of the genes in the population was determined by the responses of the listener. In each iteration, six sentences were randomly selected from the IEEE sentence database. Each sentence was processed with one set of the vocoder parameters selected from six genes. Within one iteration, the same set of six sentences was used, but the set of sentences changed from one iteration to the other. The vocoder processing between the iterations took 10 to 20 sec using Matlab on a PC with Pentium 4 processor, 3.0 GHz CPU, and 512 MB RAM. The processed sentences were evaluated by the subjects

with paired comparisons, presented in an AB comparison scheme, as this method was shown to be a robust paradigm for similar listening tasks (Eisenberg et al., 1997; Studebaker et al., 1982). The subjects were asked to choose the sentence in the pair with higher intelligibility. An option was also available to indicate that the pair was equally intelligible (or equally nonintelligible). The subjects could play the pair of sentences multiple times. No feedback was provided. With six genes in the population, there were 15 paired comparisons to complete for each iteration. The genes that were selected to have higher intelligibility by the subject had higher fitness values. The genes were then rank-ordered such that the genes with the highest and lowest fitness were ranked as the top and the bottom genes, respectively. The next generation of genes was produced from the rank-ordered genes of the old population using the methods mentioned above. Using elitism, the top two genes from the old population passed to the next generation without any change. The third gene was also passed to next generation, but this gene had a probability of being mutated. The fourth and fifth genes were generated with cross-over; two pairs of parent genes were randomly selected from the old population, with a uniform probability distribution, and the two offspring genes were created by averaging the parameters from the parent genes. These genes also had a probability of being mutated. For mutation, two of the three genes (the third, fourth, and fifth genes of the new population) were randomly selected. One randomly selected parameter of each of the two genes was changed to a randomly selected value. The sixth gene in the old population was not used in producing the next generation of genes; it was simply discarded and the sixth gene in the new population was produced randomly. The main purpose of the sixth gene was to increase the diversity of the genes in the new population. The implementation of the GA in the present study promoted diversity in the gene population except for the top two genes.

During the data collection, speech intelligibility was not measured. Instead, the subjects were asked to judge the intelligibility of the sentences that were presented and to indicate a preference for the sentence that sounded more intelligible. Because the only input to the program was user preference, there was no analytical expression for the error and a conventional convergence criterion that is based on minimization of the error could not be used. As an alternative, a new convergence criterion was defined: if the same two genes were ranked as the best genes in the population in three consecutive iterations, it was assumed that a good solution was found and the GA program was stopped. If the GA failed to converge in 15 iterations, then the program was stopped manually. The gene that was ranked as the top gene in the final iteration was accepted as the final solution.

Experimental Procedures

The processing of the speech materials and the programming of the GA were done by using Matlab. Subjects were tested in a sound-proof booth with stimuli presented binaurally over Sennheiser HD 580 headphones at a comfortable level of 65 dB SPL. Matlab GUI tools and a TDT System III were used for the presentation of the stimuli and for collecting the input from the subjects.

In the practice session, subjects were asked to listen to a sentence processed with the noiseband vocoder and to repeat the sentence to the experimenter. They were allowed to listen to the processed sentence as many times as needed. Once the subjects repeated the sentence, they were given feedback by playing the unprocessed sentence. The processing parameters were chosen such that the sentences had good quality and high intelligibility in the beginning, and became more difficult to understand toward the end. No scores were measured in the practice session, as the purpose of the task was to familiarize the subjects with the noiseband vocoder processing and to minimize learning effects during the GA runs and speech recognition tests. The subjects practiced with 100 sentences on average during a period of half an hour.

Every subject was tested 3 times with the GA to explore the repeatability of the results. A log file was kept for the off-line analysis of the decisions the subject made in the paired comparisons and how the genes in the population evolved accordingly. The solutions produced by the GA were objectively evaluated with speech recognition tests. In a final validation test, subjects compared the solutions produced by different GA runs to each other, and the best preferred GA solution to the theoretical best solution of [20 0 0], using the same paired comparison technique. The results of this test were used to compare the objective and subjective measures of intelligibility for each subject.

In a separate GA run, the effect of the selection of the initial population was explored. In this run, the initial population was generated by using the top and the second top solutions of the three runs of the GA, rather than a randomly produced initial population. The purpose was to test if the GA would produce better solutions and/or if it would converge faster when the initial population was already close to the optimal solution.

Speech Recognition with Vocoder Parameters

The effects of changing vocoder parameters on speech recognition were shown in previous studies (Başkent, 2006; Başkent & Shannon, 2003, 2007; Friesen et al., 2001; Fu & Shannon, 1999). To observe the effects specific to this experiment, and also to ensure that the theoretical best solution of [20 0 0] produced the best performance, speech intelligibility was measured with IEEE sentences as a function of the vocoder parameters used in the present study. The stimuli were presented to the subject in a laboratory setting similar to the one used for the GA, except the subject did not see the monitor and the task was to repeat the sentence that was presented. For each condition, a set of 10 IEEE sentences was selected randomly, but different sets of sentences were used for each condition with the same subject. Each sentence was presented once and no feedback was provided. The experimenter entered the number of words heard correctly for each sentence, and the percent correct scores for each condition were determined by counting the number of correct words for all 10 sentences presented. The order of the conditions was randomized.

The percent correct scores, averaged across subjects, are shown in Figure 2. The panels from left to right show the scores as a function of the number of vocoder channels, frequency shift, and frequency compression-expansion, respectively. The scores show that best performance was observed around the largest number of channels, 20, and the smallest amount of spectral distortions, 0, as expected. However, the scores also show that there is a tolerance range around these values over which high performance was observed. This implies that there was not a single best solution, but a range of best solutions around [20 0 0].

The final settings produced by the GA were objectively evaluated by measuring speech intelligibility with the same method described above, to explore if the scores for the settings that the GA converged to were near the scores obtained with the best solution of [20 0 0].

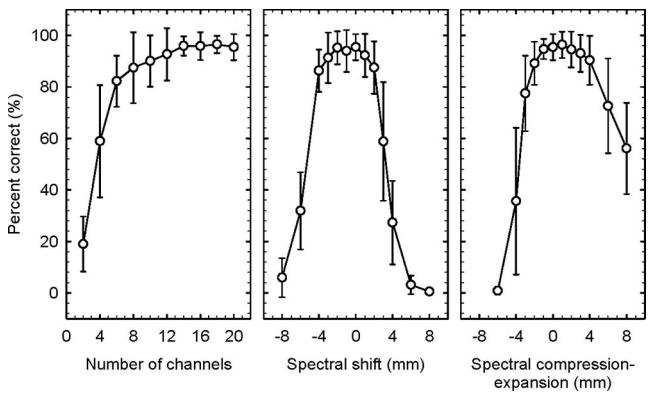

Fig. 2. Speech recognition performance, averaged across subjects, shown for each vocoder parameter separately. Error bars show ±1 standard deviation.

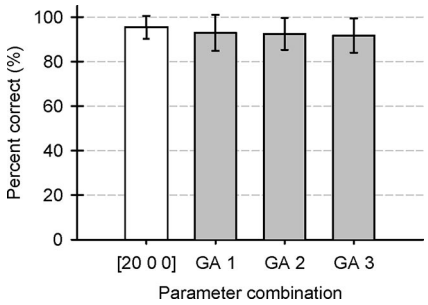

Fig. 3. Sentence recognition scores, averaged across subjects. White bar shows the average score for the baseline condition of [20 0 0]; gray bars show the average scores for the settings produced by GA runs 1 through 3. Error bars show ±1 standard deviation.

RESULTS AND DISCUSSION

Figure 3 presents the percent correct scores averaged across subjects. The white bar shows the scores measured with the theoretical best solution of [20 0 0] and the gray bars show the scores with the solutions produced by three GA runs. The average scores were similar across different GA runs. The average score produced by each GA run was around 3% lower than the average baseline score with [20 0 0]. A two-tailed paired *t*-test, applied after the scores were transformed into rationalized arcsine units (RAU; Sherbecoe & Studebaker, 2003), showed that the difference between the average score from each GA run and the baseline score with [20 0 0] was not statistically significant [$t(8) = 0.81$ and $p = 0.44$ for GA run 1, $t(8) = 1.08$ and $p = 0.31$ for GA run 2, and $t(8) = 1.49$ and $p = 0.18$ for GA run 3].

Analysis of Individual Scores

Figure 3 shows that the performance with the settings produced by the GA was similar to the performance produced with [20 0 0] when average scores were compared. The variance in the scores, however, was higher with the GA scores. The lowest score with the baseline condition of [20 0 0] was 85% while the lowest score with the GA solutions was 76%, which implied that a few GA settings probably produced poorer than optimal performance. An example of such variability in the solutions produced by the GA is shown in Table 2 for two subjects. The table shows that the GA did not always produce the same settings, but generally the intelligibility scores measured with the settings from different GA runs

TABLE 2. Examples of performance with the best theoretical solution of [20 0 0] and with solutions produced by the individual GA runs, shown for subjects S2 and S12

Subject S2			
Settings	Percent correct score (%)	Number of iterations	Convergence time (min)
Baseline [20 0 0]	86	—	—
GA run 1 [20 −1 0]	85	15	47
GA run 2 [17 0 −1]	93	9	30
GA run 3 [16 2 1]	90	12	45
Subject S12			
Settings	Percent correct score (%)	Number of iterations	Convergence time (min)
Baseline [20 0 0]	100	—	—
GA run 1 [20 −2 1]	100	6	16
GA run 2 [18 1 2]	100	7	17
GA run 3 [18 2 8]	84	6	17

were comparable to each other and also comparable to the performance measured with the baseline setting of [20 0 0]. However, the third run of the GA produced a solution that resulted in poorer speech recognition with subject S12.

To explore the variability in the solutions produced by the GA runs, scores were analyzed for individual subjects. Figure 4 shows the differences between the percent correct scores produced by the GA settings and the score produced with [20 0 0] for each subject and for each GA run. A negative score indicates that the GA setting resulted in poorer performance. The GA scores were compared with each subject's own score with [20 0 0], rather than the score averaged across subjects. The scores for the baseline condition of [20 0 0] were low (i.e., more than one

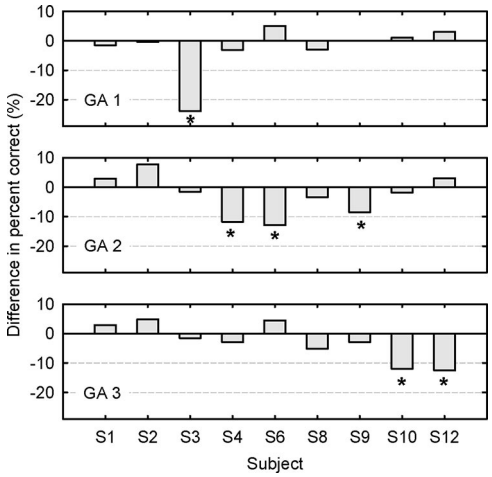

Fig. 4. The difference between the scores produced with the GA settings and the score produced with the baseline setting of [20 0 0], shown for each subject individually. Panels from top to bottom show the results for GA runs 1 through 3, respectively.

standard deviation below the average score) for two subjects. Varying linguistic skills across subjects might have caused such variation in the baseline scores. For example, some subjects were raised in bilingual families even though they used English as the primary language starting from young ages, and such factors have been shown to affect speech recognition in challenging listening environments (Rogers et al., 2006). An example of this situation is subject S2, who was raised speaking Cantonese and English. This subject's baseline score was considerably lower than for the other NH subjects, as shown in Table 2.

The bars denoted with a star show GA scores that were lower than the 95% confidence interval of the subject's baseline score. The confidence interval was calculated using the method suggested by Thornton & Raffin (1978) for open-set word recognition tests. The method assumes that stimuli are of equal difficulty and the responses are independent, and the subject's responses can be modeled as binomial distribution. When presented in sentences, the probability of correct identification of words is usually higher than for words presented in isolation due to context effects (Olsen et al., 1997). For the sentences used in the present study, however, the context effects were shown to be relatively low (Rabinowitz et al., 1992), and this factor was not included in the calculations of the confidence interval.

Figure 4 shows that many subjects performed similarly (i.e., a difference less than 5%) with the solution produced by the GA and with the baseline setting of [20 0 0]. However, in each GA run, a few GA solutions were observed to produce poorer results. When the results from GA run 1 and GA run 2 (top and middle panels in Figure 4) were considered together, it was observed that there was at least one good solution produced by the GA for each subject.

Convergence

The intelligibility scores and the convergence properties were similar across the three GA runs. Therefore, the data from GA runs 1 through 3 were pooled across subjects and across runs, to calculate the average number of iterations and the average convergence time. The average number of iterations was 8.0, with a standard deviation of 3.6, and the average convergence time was 25.5 minutes, with a standard deviation of 13.2 minutes.

There was a variation in the convergence properties among subjects. Convergence for subject S2, for example, was slower than for the other subjects (Table 2). However, the GA produced good solutions in each run with this subject. On the other hand, the second run of the GA converged in three iterations

with subject S6, but produced a poor final result (Figure 4). In this run, the GA did not produce better solutions in the first iterations and the subject had to indicate preferences for the best solution in a population of poor solutions. Hence, the GA stopped prematurely, following the convergence criterion, before having a chance to produce good solutions.

Objective Versus Subjective Judgment of Intelligibility

In the present study, the only input to the GA from the subjects was the subjective judgment of the intelligibility of the sentences presented. The GA was able to produce solutions with good intelligibility, which suggests that the subjects were sufficiently successful in providing reliable subjective input to the GA program.

After the GA runs, two tests were conducted; one to compare the solutions produced by different GA runs to each other, and one to compare the best preferred GA solution with the theoretical best solution of [20 0 0]. In the first test, subjects compared their own solutions from the 3 runs of the GA with paired comparisons. Figure 5 shows the difference between the scores for the GA solutions and for [20 0 0], as in Figure 4, except that all three runs are shown next to each other for each subject in the same panel. The numbers above and below the bars show the preferences of the subjects, as determined by the paired comparison test. A solution denoted by "1" was ranked as the best solution among all GA produced solutions. As a result, the figure shows the quality of the GA solutions both with an objective measure, shown by percent correct scores, and a subjective measure, shown by subject preferences, of intelligibility. For many subjects, such as S1, S2, S3, S6, S9, and S12, there was good agreement between the two measures. For two subjects, S4 and S10, however, the two measures did not match. The

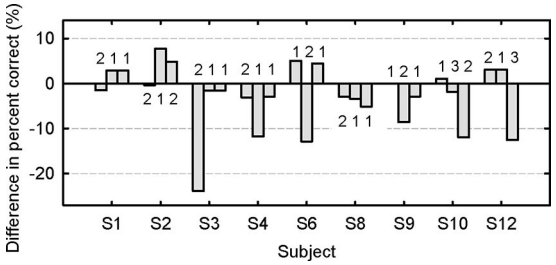

Fig. 5. Difference between the scores produced by the GA settings and the score produced by the baseline setting of [20 0 0], shown for each subject individually. Different than Figure 4, the scores are shown next to each other for GA runs 1 through 3 for each subject. Numbers above and below the bars show the subjective ranking of the GA solutions according to subject preferences.

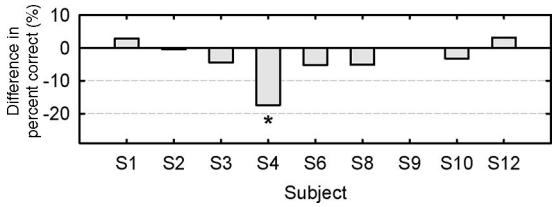

Fig. 6. Difference between the scores produced by the GA that started with a better initial population and the score produced by the baseline setting of [20 0 0].

second test, where the subjects compared their best-ranked GA solution to the best theoretical solution of [20 0 0], showed that all nine subjects had equally strong preference for the best-ranked GA solution and the [20 0 0] setting.

Effect of Initial Population

In a separate run of the GA, the best and second best solutions from the three GA runs were used to generate the initial population. The purpose was to observe if an initial population that included good settings, rather than randomly generated ones, would result in solutions with higher intelligibility and/or faster convergence.

The average score from this run was 92.1%, with a standard deviation of 7.2%. This score was comparable to the average scores observed in the GA runs that started with a random population (Figure 2). Figure 6 shows the difference in the scores between the solution produced by the GA that started with a better initial population and the baseline condition of [20 0 0] for each subject. For most subjects the scores were similar. Only for S4 was the percent correct score measured with the solution produced by this GA lower than the 95% confidence interval, calculated around the baseline condition.

The GA that started with a better population produced similar solutions to the GAs that started with randomly selected populations; however, a benefit was observed in the speed of convergence. The GA that started with a better population converged in 5.3 iterations on average, with a standard deviation of 3.8, and in 17.9 minutes on average, with a standard deviation of 12.2 minutes. Both were significantly lower, as shown by a paired t-test ($p < 0.05$), than the average number of iterations (8.0) and the average convergence time (25.5 minutes) observed for the three GA runs that started with a random population of genes.

General Discussion

Previous studies have applied optimization algorithms to fitting hearing aid or cochlear implant

settings with actual device users and the results showed potential for these algorithms. However, since there was not a good understanding of the nature of all possible solutions for all subjects, it was not easy to objectively evaluate the final settings, especially with respect to all possible solutions. The reliability of the modified simplex algorithm, suggested for fitting gain with hearing-impaired listeners, was often evaluated by subjective judgment of the settings produced by the algorithm (Kuk & Pape, 1992; Neuman et al., 1987; Preminger et al., 2000), or by comparing these settings with the most commonly used prescriptive gain methods (Preminger et al., 2000). Test-retest reliability was explored by running the algorithm multiple times and comparing the settings produced in each run (Kuk & Pape, 1992; Stelmachowicz et al. 1994). Preminger et al. (2000) also measured speech recognition with the gain settings produced by the simplex algorithm. The scores were similar to those measured with the NAL-R prescription. Durant et al. (2004) used the GA to optimize feedback cancellation with hearing-impaired listeners. The final settings were evaluated by measuring the gain margin, that is, the maximum gain that could be applied before feedback occurred. There was no subjective or objective evaluation by subjects. Wakefield et al. (2005) used the GA to optimize many cochlear implant parameters simultaneously with implant users. Speech recognition performance was similar for the settings produced by the GA and the settings of the patient's own device. There was no data for subjective preference by the subjects.

The present study presented a fully controlled listening problem with many possible solutions where the best and worst solutions were known and where the solution produced by the GA could be evaluated objectively. The results, therefore, are complementary to the studies mentioned above, and provide a theoretical frame work for studies that involve more realistic listening conditions with real patients.

The results showed that speech recognition performance measured with the vocoder settings produced by the GA was, on average, comparable to the performance measured with the best setting of [20 0 0]. When the data were analyzed for individual subjects and for each GA run, occasionally there was a solution that produced poorer performance. Several factors might have contributed to the occasional poor solution produced by the GA: inherent randomness of the GA, the specific convergence criterion used in the present study, and possible mistakes or misjudgments that the users might have made in the paired comparisons. For practical applications, it might be more beneficial not to use such an automatic stopping criterion, but to allow the GA to run for a specified number of iterations or a certain amount of time so that the GA could continue searching for better settings. Another alternative would be to run the GA twice, as it was observed that there was at least one good solution produced for each subject when the results were combined for the first two runs of the GA. The two solutions produced by two GA runs, for example, can be programmed into two memories of the hearing aid or cochlear implant, so that the patient can have an opportunity to evaluate both settings for an extended time and for diverse listening conditions.

With the current implementation of the GA, the average convergence time was around 25 minutes. This time frame suggests that the GA is faster than an exhaustive search where the listeners would have to evaluate all possible settings. In the present study, for example, the step sizes and the lower and upper limits selected for the parameters resulted in 4845 possible solutions. With the advantage of being able to find a good solution among thousands of possible solutions in less than half an hour, the GA could be a useful tool in research for finding the best settings for a new feature under development. However, for clinical applications, the running time might have to be shortened. In the present study, the main objective was to show that the GA can work with subjective input from listeners. Therefore, the particular implementation of the GA used in the study was not optimized for practical applications. For example, there were six genes in the population and all were compared to each other in each iteration, producing 15 paired comparisons. A major portion of the average running time of 25 minutes was used for these comparisons. A possible alternative would be to use five genes, rather than six, which could still be a reasonable size for an effective algorithm, whereas the number of paired comparisons is instantly reduced to 10. Moreover, it is probably redundant to compare all possible pairs of genes in the population, as the outcome of some comparisons can be deduced from previous comparisons using transitivity properties, assuming that the subject provides consistent input. For a real-life application one can take advantage of this redundancy to reduce the number of paired comparisons. Another potential improvement might be to start the population with settings that are estimated to be close to the best setting, which was observed to result in faster convergence.

Since the overall performance measured with the solutions produced by the GA was high, the subjects must have been able to provide reliable input to the GA. When the subjective preferences were compared with the objective measures, there was a good agree-

ment in general except for a small number of subjects, similar to findings of Punch & Parker (1981). The particular implementation of the GA used in the present study seems to be able to produce good results even if there were a few inconsistencies in the paired comparisons, most probably because all genes were compared to each other in every iteration. If a smaller number of comparisons are made, as suggested in the previous paragraph to shorten the running time, and the rest are inferred from previous comparisons, such inconsistencies might carry over to following iterations and might cause the GA to produce poor solutions.

One way of ensuring reliability of user input would be to use the GA with a listening problem where the processing changes speech intelligibility or sound quality sufficiently, such that the user can perceive the differences in the sounds presented in pair comparisons and so can form a judgment. For this reason, the tasks of fitting linear gain or compression, which produce subtle effects, might not be the most ideal problems for the GA. For these tasks, there are already many methods available; the conventional prescriptive procedures, such as NAL-R, NAL-NL1, and POGO, provide a fast fitting, while adaptive procedures, such as Scaladapt, Camadapt, and Ear Tuner, provide customization. The strength of the GA, especially due to its capability of optimizing multiple parameter combinations simultaneously, seems to be in finding the best settings for more complex features, where a large number of possible solutions have to be evaluated.

The main purpose of the present study was to show that the GA can work in ideal laboratory settings before it can be suggested for real-life applications. Therefore, an ideal group of subjects, young (between the ages of 19 and 34 yr) and healthy with normal hearing, was recruited. The performance of this subject group can be considered as an ideal case. There will be many elderly hearing aid or cochlear implant users who might have additional difficulties in listening tasks due to perceptual or cognitive deficiencies. With such subjects, the overall performance and efficiency of a GA program might be lower than found in the present study.

CONCLUSION

The results showed that human subjects were generally able to provide sufficiently reliable subjective input to the GA to produce solutions with good intelligibility, which suggests that the GA has potential for real-life applications such as optimization of device settings of hearing aids or cochlear implants. The GA implementation used in the present study would be a fast method for finding optimal settings for new device features under research. However, for clinical applications, the program might need to be modified to be more practical, without compromising on reliability.

ACKNOWLEDGMENTS

The authors would like to thank Qian-Jie Fu and John Galvin for permission to use IEEE sentences recorded in the Speech Technology and Hearing Research Lab at House Ear Institute, Robert V. Shannon for constructive discussions, and Brian C. J. Moore and an anonymous reviewer for their comments on an earlier version of this paper.

Address for correspondence: Dr. Deniz Başkent, Starkey Hearing Research Center, 2150 Shattuck Avenue, Ste. 408, Berkeley, CA 94704. E-mail: deniz_baskent@starkey.com.

Received January 13, 2006; accepted December 21, 2006.

REFERENCES

Allen, J., Hall, J., Jeng, P. (1990). Loudness growth in ½-octave bands (LGOB) - A procedure for the assessment of loudness. *Journal of the Acoustical Society of America*, *88*, 745–753.

Başkent, D. (2006). Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels. *Journal of the Acoustical Society of America*, *120*, 2908–2925.

Başkent, D., Shannon, R. V. (2003). Speech recognition under conditions of frequency-place compression and expansion. *Journal of the Acoustical Society of America*, *113*, 2064–2076.

Başkent, D., Shannon, R. V. (2007). Combined effects of frequency-place compression-expansion and shift on speech recognition. *Ear and Hearing*, *28*, 277–289.

Berger, R. A., Hagberg, E. N., Rane, R. L. (1977). Prescription of hearing aids: rationale, procedures, and results. Kent, Ohio: Herald Publishing House.

Bourgeois-République, C., Chabrier, J. J., Collet, P. (2005). An interactive evolutionary algorithm for cochlear implant fitting: first results. Proc. 2004 ACM Symp. App. Comp., 231-235, Santa Fe, New Mexico.

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

Byrne, D., Cotton, S. (1988). Evaluation of the National Acoustic Laboratories new hearing aid selection procedure. *Journal of Speech, Language, and Hearing Research*, *31*, 178–186.

Byrne, D., Dillon, H., Ching, T., Katsch, R., Keidser, G. (2001). NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures. *Journal of the American Academy of Audiology*, *12*, 37–51.

Dudley, H. (1939). Remaking speech. *Journal of the Accoustical Society of America*, *11*, 169–177.

Durant, E. A., Wakefield, G. H., Van Tasell, D., Rickert, M. E. (2004). Efficient perceptual tuning of hearing aids with genetic algorithms. *IEEE Transactions on Speech and Audio Processing*, *12*, 144–155.

Eisenberg, L. S., Dirks D. D., Gornbein, J. A. (1997). Subjective judgements of speech clarity measured by paired comparisons and category rating. *Ear and Hearing*, *18*, 294–306.

Franck, B. A. M., Dreschler, W. A., Lyzenga, J. (2004). Methodological aspects of an adaptive multidirectional pattern search to optimize speech perception using three hearing-aid algorithms. *Journal of the Acoustical Society of America*, *116*, 3620–3628.

Friesen, L. M., Shannon, R. V., Başkent, D., Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America*, *110*, 1150–1163.

Fu, Q.-J., Shannon, R. V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *Journal of the Acoustical Society of America*, *105*, 1889–1900.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. (1993). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354.

Green, T., Faulkner, A., Rosen, S., Macherey, O. (2005). Enhancement of temporal periodicity cues in cochlear implants: effects on prosodic perception and vowel identification. *Journal of the Acoustical Society of America*, *118*, 375–385.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species: 29 years later. *Journal of the Acoustical Society of America*, *87*, 2592–2605.

Institute of Electrical and Electronics Engineers (1969). IEEE recommended practice for speech quality measurements.

Kiessling, J., Schubert, M., Archut, A. (1996). Adaptive fitting of hearing instruments by category loudness scaling (ScalAdapt). *Scandinavian Audiology*, *25*, 153–160.

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

McCandless, G. A., Lyregaard, P. E. (1983). Prescription of gain/output (POGO) for hearing aids. *Hearing Instruments*, *34*, 16–21.

McDermott, H. (1998). A programmable sound processor for advanced hearing aid research. *IEEE Transactions on Rehabilitation Engineering*, *6*, 53–59.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill, International Ed.

Moore, B. C. J., Marriage, J. E., Alcantara, J. I., Glasberg, B. R. (2005). Comparison of two adaptive procedures for fitting a multi-channel compression hearing aid. *International Journal of Audiology, 44*, 345–357.

Neuman, A. C., Levitt, H., Mills, R., Schwander, T. (1987). An evaluation of three adaptive hearing aid selection strategies. *Journal of the Acoustical Society of America*, *82*, 1967–1976.

Olsen, W. O., Van Tasell, D. J., Speaks, C. E. (1997).Phoneme and word recognition for words in isolation and in sentences: the Carhart Memorial Lecture, American Auditory Society, Salt Lake City, Utah 1996. *Ear and Hearing, 18*, 175–188.

Poissant, S. F., Whitmal, N. A., Freyman, R. L. (2006). Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *Journal of the Acoustical Society of America*, *119*, 1606–1615.

Preminger, J. E., Van Tasell, D. J. (1995). Measurement of speech quality as a tool to optimize the fitting of a hearing aid. *Journal of Speech and Hearing Research*, *38*, 726–736.

Preminger, J. E., Neuman, A. C., Bakke, M. H., Walters, D., Levitt, H. (2000). An examination of the practicality of the simplex procedure. *Ear and Hearing*, *21*, 177–193.

Punch, J. L., Parker, C. A. (1981). Pairwise listener preferences in hearing aid evaluation. *Journal of Speech and Hearing Research*, *24*, 366–374.

Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., Cuneo, P. A. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *Journal of the Acoustical Society of America*, *92*, 1869–1881.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., Abrams, H. M. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, *27*, 465–485.

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

Sherbecoe, R. L., Studebaker, G. A. (2003). Supplementary formulas and tables for calculating and interconverting speech recognition scores into transformed arcsine units. *International Journal of Audiology*, *43*, 442–448.

Stelmachowicz, P. G., Lewis, D. E., Carney, E. (1994). Preferred hearing aid frequency responses in simulated listening environments. *Journal of Speech and Hearing Research*, *37*, 712–719.

Studebaker, G. A., Bisset, J. D., Van Ort, D. M., Hoffnung, S. (1982). Paired comparison judgments of relative intelligibility in noise. *Journal of the Acoustical Society of America*, *72*, 80–92.

Takayagi, H. (2001). Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. *Proceeding of the IEEE*, *89*, 1275–1296.

Thornton, A. R., Raffin, M. J. (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research*, *21*, 507–518.

Valente, M., Van Vliet, D. (1997). The independent hearing aid fitting forum (IHAFF) protocol. *Trends of Amplification* , *2*, 6–35.

Wakefield, G. H., van den Honert, C., Parkinson, W., Lineaweaver, S. (2005). Genetic algorithms for adaptive psychophysical procedures: recipient-directed design of speech-processor MAPs. *Ear and Hearing*, *26*:57S–72S.

Xu, L., Thompson, C. S., Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *Journal of the Acoustical Society of America*, *117*, 3255–3267.

Zakis, J. A. (2003). A trainable hearing aid. PhD Thesis, Department of Otolaryngology, University of Melbourne, Victoria, Australia.