Simulating listener errors in using genetic algorithms for perceptual optimization

Deniz Başkent and Brent Edwards
Starkey Hearing Research Center, 2150 Shattuck Ave., Ste. 408, Berkeley, California 94704
deniz_baskent@starkey.com, brent_edwards@starkey.com

Abstract: The genetic algorithm (GA) was previously suggested for fitting hearing aid or cochlear implant features by using listener's subjective judgment. In the present study, two human factors that might affect the outcome of the GA when used for perceptual optimization were explored with simulations. Listeners with varying sensitivity in discriminating sentences of different intelligibility and with varying error rates in entering their judgment to the GA were simulated. A comparison of the simulation results with the results from human subjects reported by Başkent *et al.* Ear Hear. **28**(3) 277–289 (2007) showed that these factors could reduce the performance of the GA considerably.

PACS numbers: 43.66.Yw, 43.71.An, 43.66.Ts [QJF]
Date Received: January 4, 2007 **Date Accepted:** March 22, 2007

1. Introduction

Most modem hearing aids and cochlear implants offer many features in addition to providing basic audibility. As individual users might have different pathologies and listening preferences (Preminger and Van Tasell, 1995), the numerous device features need to be customized for each patient to maximize benefit. This adjustment can be a complicated and time-consuming process, especially if some of the device features also interact with each other.

Optimization algorithms can be used as a tool to achieve the individual customization in a reasonable time. The modified simplex algorithm was proposed for fitting gain in hearing aids (Kuk and Pape, 1992; Neuman *et al.*, 1987; Preminger *et al.*, 2000; Stelmachowicz *et al.*, 1994). Genetic algorithms (GAs) were suggested for fitting features related to hearing aids (Durant *et al.*, 2004) or cochlear implants (Bourgeois-République *et al.*, 2005; Wakefield *et al.*, 2005). In such perceptual optimization, candidate parameter sets are first evaluated by a listener and then modified according to listener's preferences following the rules of the particular method used. The steps of evaluation and modification continue iteratively, until a satisfactory set of parameters is found. The main advantages that optimization algorithms offer are speed, because the final optimal solution is typically reached by evaluating only a fraction of all possible solutions, and flexibility, because they can be implemented to optimize any device feature.

In perceptual optimization, the input to the program is the subjective human response and the appropriateness of the final solution is, again, judged by the listener. Therefore, there is often no metric available to quantitatively analyze how well the program works (Takayagi, 2001). Başkent *et al.* (2007b) systematically distorted speech using three parameters of the noiseband vocoder processing (Shannon *et al.*, 1995), to generate a listening problem with a metric. The acute effects of these manipulations on intelligibility of speech by normal-hearing subjects were known from previous studies (Fu and Shannon, 1999; Başkent and Shannon, 2003; 2007a; Başkent, 2006), so the final solutions produced by the GA could similarly be evaluated. Speech intelligibility scores measured with the settings produced by the GA were, on average, very high, indicating that the subjects must have been able to provide sufficiently reliable subjective input. Analysis of data from individual subjects showed that there was generally a good agreement between the subjective and objective measures of intelligibility, and only a small number of inconsistencies were observed.

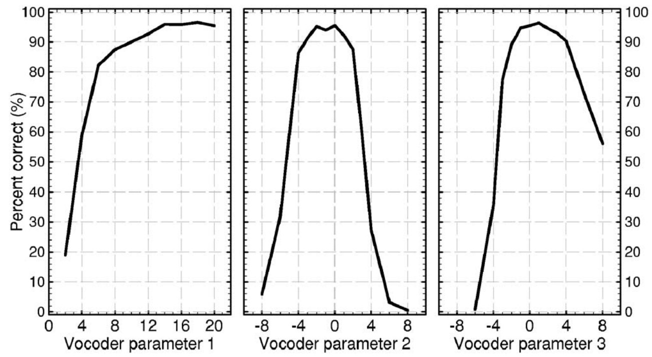

Fig. 1. Speech recognition performance, averaged across nine normal-hearing subjects, shown for each vocoder parameter separately. Reproduced from Başkent *et al.* (2007b).

The subjects who participated in the study by Başkent *et al.* (2007b) were young, with no auditory or cognitive deficits. Therefore, the results can be interpreted as how well the GA would work with ideal listeners. In real applications, some hearing aid and cochlear implant users might have difficulty in making a reliable judgment due to varying peripheral or central auditory deficits or diminished cognitive skills, for example, as a result of aging. In the present study, the effects of such human factors on perceptual optimization with the GA are explored with simulations. One factor that was simulated was the sensitivity in distinguishing sentences of varying intelligibility. The second factor was the errors a subject might make in entering the subjective input into the GA. The same GA program was used as Başkent *et al.* (2007b) study, and the results from the simulations were compared to the results with real listeners, reported in the same study.

2. Methods

2.1 Noiseband vocoder processing

Noiseband vocoder has been widely used to systematically explore the effects of temporal and spectral degradations on speech perception, or to simulate cochlear implant processing with normal-hearing subjects. Narrow bands of noise (carrier bands) are modulated with envelopes extracted from individual bands of speech (analysis bands). The processed speech, a synthesis of these modulated noise bands, has only the crude spectral and temporal elements of the input speech (Shannon *et al.*, 1995).

Başkent *et al.* (2007b) had selected three vocoder parameters to optimize with the GA: (1) the number of the spectral channels of the vocoder, (2) a shift between the analysis and carrier band frequency ranges, and (3) a widening/narrowing of analysis band frequency range over the carrier band frequency range. The percent correct scores with IEEE sentences (IEEE, 1969), averaged across nine normal-hearing subjects, are reproduced from Başkent *et al.* (2007b) in Fig. 1 for each of the three parameters.

The intelligibility of a solution produced by the GA with a simulated subject was evaluated with predicted percent correct (PPC), a measure estimated from a multiplicative combination of the average scores, shown in Fig. 1 for each vocoder parameter. Hence, the effects of the three vocoder parameters were assumed to be independent, even though Başkent and Shannon (2007a) had shown that there was interaction between vocoder parameters 2 and 3 for a small number of conditions.

2.2 Genetic algorithm

For consistency, the same GA that was used by Başkent *et al.* (2007b) was implemented in the present study. The GA is an inherently stochastic optimization method that is based on concepts related to the evolution theory (Mitchell, 1997). For example, one set of parameters that will be optimized is called a gene. In the present study, every gene was a combination of the three vocoder parameters mentioned in the previous section. The levels of the parameters 1 to 3 were selected as 19, 17, and 15, respectively, producing a search space of 4845 possible solutions. Unlike the conventional bitstring coding, actual parameter values were used in the genes. GAs work on a population of genes (six was used in the present study) rather than an individual set of parameters, and the genes in the initial population are generated randomly. In the present study, a uniform distribution was used for all random processors, except for the mutation operator. In each iteration, all genes in the population are evaluated for fitness and genes with better fitness have a higher probability to pass to the next generation. In applications that involve human subjects, the fitness is determined by the listener's preferences. Başkent *et al.* (2007b) presented vocoder-processed IEEE sentences in paired comparisons, 15 to compare all six genes to each other, to the subjects. The subjects were asked to enter a preference for the sentence with higher subjective intelligibility (A better than B, or vice versa), with an additional option for equal intelligibility (A B same). The genes that were preferred more often had higher fitness value, and all six genes of the population 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 one of these methods: (1) Elitism: the top two genes with highest fitness values passed onto the next generation with no alterations. The top third gene was also passed onto the next generation, but with a probability of being mutated. (2) Cross-over: two non-identical parent genes were randomly selected from the old population, and two new child genes were produced by averaging the parameters from the parent genes. The offspring genes replaced the fourth and fifth genes of the old population. (3) Mutation: two of the three genes (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, using a normal distribution with the mean at the parameter's old value and the standard deviation of one third of the number of levels used for the parameter to be mutated. The sixth gene in the old population was not used in producing the next generation of genes; the old one was discarded and the sixth gene of the new population was produced randomly. These steps were repeated iteratively, until a convergence criterion was satisfied: if the same two genes were ranked as the best genes of the population in three consecutive iterations, convergence was assumed. 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 optimal solution.

2.3 Simulations

Başkent *et al.* (2007b) compared objective and subjective measures of intelligibility and in a small number of occasions subjects were not accurate in judging the intelligibility of a sentence. If this happens during the comparison of a pair of sentences, the subject might enter a higher preference for the sentence with lower intelligibility that might lead the GA toward poorer solutions. This factor was modeled by the probability of error (P_{err}), the probability of making an incorrect decision in a paired comparison. For small values of P_{err}, the simulated listener makes fewer mistakes in selecting the sentence with higher intelligibility in the paired comparisons. For very high values of P_{err}, the simulated subject frequently enters incorrect choices, leading the GA to produce poorer solutions.

A second factor that could affect the outcome of the GA would be the just noticeable difference (JND) between the intelligibility levels of the sentences presented in a pair. The subject has to be able to hear the difference between the sentences to make a judgment, and how much of a difference a subject needs for a reasonable judgment most likely varies from subject to subject. This factor was modeled with the parameter $\mathrm{JND}_{\mathrm{PC}}$. In the simulations, the intelligibility related to a set of vocoder parameters was directly estimated by the PPC. The simulated subject entered a preference for one of the sentences in the pair only if the absolute difference in the intelligibility of the sentences, expressed in PPC scores, was larger than $\mathrm{JND}_{\mathrm{PC}}$. Otherwise, there was no preference and "A B same" option was selected. A small $\mathrm{JND}_{\mathrm{PC}}$ models a subject

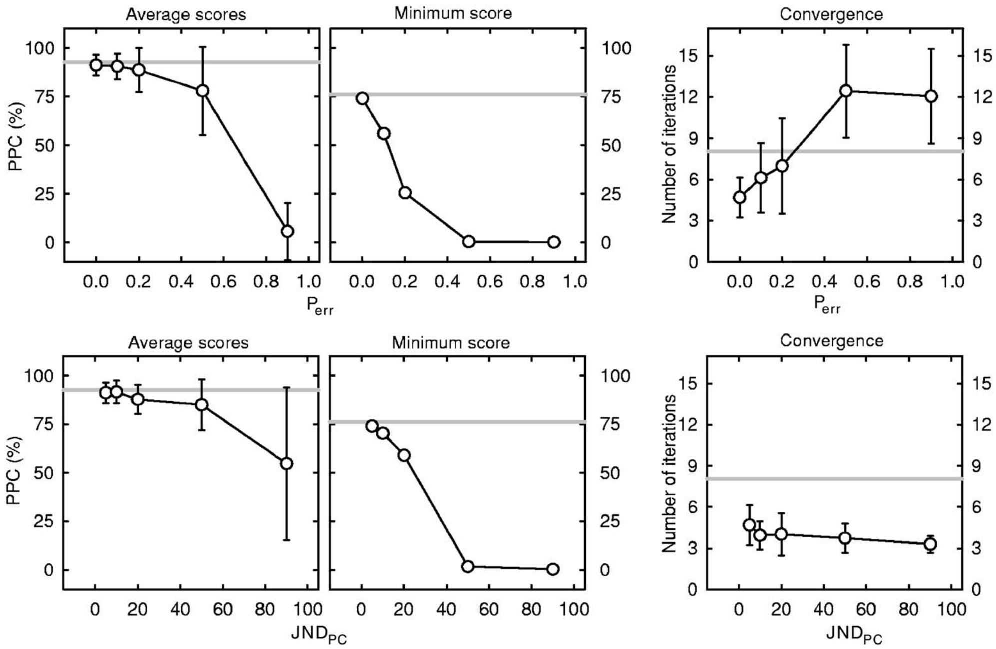

Fig. 2. Simulation results, averaged from 50 GA runs. The upper row shows the simulated performance as a function of P_{err}, probability of error in paired comparisons, and the lower row shows the performance as a function of the JND_{PC}, the smallest difference in percent correct scores that the simulated subject can perceive between the intelligibility levels of two sentences. In each row, the panels from left to right show the average predicted percent correct (PPC) scores, the lowest PPC score observed in 50 GA runs, and the average number of iterations. The gray lines show the data by real listeners, adapted from Başkent *et al.* (2007b). The error bars show one standard deviation.

that can hear a small difference and can correctly judge which sentence is more intelligible. For higher values of JND_{PC}, the simulated subject registers more of the "A B same" option. Therefore, this factor does not produce an error *per se*; rather it decreases the amount of useful information entered into the GA.

3. Results

The effects of the simulated factors were explored by running multiple simulations of the GA and observing the changes in the overall performance. Figure 2 presents the results for the factors of P_{err} and JND_{PC} in the upper and lower panels, respectively. The figure shows the effects of each factor individually; when P_{err} was varied, JND_{PC} was equal to 5%, and when JND_{PC} was varied, P_{err} was equal to 0. The panels from left to right show the average PPC scores averaged across 50 runs, the minimum PPC score of the 50 runs, and the number of iterations at convergence averaged across 50 runs. The gray lines show the corresponding data with real subjects, adapted from Başkent *et al.* (2007b). In each panel, the smallest values of P_{err} and JND_{PC}, 0 and 5%, respectively, simulated the ideal listener. The performance remained high for $P_{err} \leqslant 0.10$ and $JND_{PC} \leqslant 10\%$. As the value of P_{err} increased, the simulated listener made more errors in the paired comparisons, and the overall performance, shown by average PPC, and the probability of the GA producing a poor result in an individual run, shown by the minimum PPC score, decreased, both reaching 0% for very large P_{err}. For high P_{err} values, the number of iterations needed for convergence also increased as the user preferences were not consistent from one iteration to the next. For $P_{err} \geqslant 0.50$, the GA failed to converge for most of the runs and was manually stopped by 15 iterations. JND_{PC} had similar effect on the average and minimum PPC scores. For large JND_{PC} values, the average PPC approached 50% as there was almost no useful information entered by the simulated subject into the GA and the GA would produce random results that are dominated by the initial random gene population. Similar convergence was observed for all $\mathrm{JND}_{\mathrm{PC}}$ values. For large $\mathrm{JND}_{\mathrm{PC}}$ values, this situation indicated a premature convergence, as the GA produced poor solutions despite the fast convergence.

Simulations showed that both factors could affect the outcome of the GA negatively. Further simulations, not included in the present manuscript to ensure brevity, showed that combined effects of these factors could lead to poorer solutions and/or convergence. The experimental data from human subjects, as shown by the gray lines, was most similar to the ideal listener, implying that the input by real subjects into the GA program was sufficiently reliable.

4. Conclusion

Başkent *et al.* (2007b) showed that the GA can produce reasonable solutions with young normal-hearing listeners under controlled laboratory settings. When the data with human listeners was compared to the data with simulated listeners of the present study, it was observed that the performance by real listeners was similar to the ideal user, who was able to distinguish sentences with a small difference in intelligibility and who was also fairly accurate with paired comparisons. Simulations also showed that the particular GA implementation by Başkent *et al.* (2007b) could handle these factors for small values, most probably because all genes were compared to each other in every iteration, which provided plenty of information and many chances for the GA to correct itself. However, for larger values, simulating a situation more likely to occur with elderly and/or hearing-impaired listeners, performance dropped considerably.

Previous studies had proposed optimization algorithms for customizing hearing aids (Durant *et al.*, 2004; Kuk and Pape, 1992; Neuman *et al.*, 1987; Preminger *et al.*, 2000; Stelmachowicz *et al.*, 1994) or cochlear implants (Bourgeois-République *et al.*, 2005; Wakefield *et al.*, 2005) with real patients. Even though the simulation results of the present study would be applicable specifically to the GA implementation reported by Başkent *et al.* (2007b), similar simulations could be used to characterize the effects of differing user skills on how well any perceptual optimization method might work for general population. For example, Başkent *et al.* (2007b) suggested that a smaller number of paired comparisons are made, with the rest being inferred from previous comparisons, to shorten the running time. However, if the listener makes many errors, these errors might carry over to following iterations, and might cause the GA to produce poorer solutions. Using the present study as a guideline, similar simulations can be developed to use as a tool for assessment of such potential modifications. A customized simulation method could be useful in evaluating the potential success of a specific optimization program and also in deciding which operators would result in best performance, in a faster manner before the actual testing with human listeners.

References and links

Başkent, D. (**2006**). "Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels," J. Acoust. Soc. Am. **120**, 2908–2925.

Başkent, D., and Shannon, R. V. (**2003**). "Speech recognition under conditions of frequency-place compression and expansion," J. Acoust. Soc. Am. **113**, 2064–2076.

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

Başkent, D., Eiler, C. N., and Edwards, B. (**2007b**). "Using genetic algorithms with subjective input from human subjects: Implications for fitting hearing aids and cochlear implants," Ear Hear. **28**, 370–380.

Bourgeois-République, C., Chabrier, J. J., and Collet, P. (**2005**). "An interactive evolutionary algorithm for cochlear implant fitting: First results," *Proceedings of the 2004 ACM Symposium on Applied Computings*, pp. 231–235, Santa Fe, New Mexico.

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

Fu, Q.-J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," J. Acoust. Soc. Am. **105**, 1889–1900.

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

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

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

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**, 1967–1976.
Preminger, J. E., and Van Tasell, D. J. (**1995**). “Measurement of speech quality as a tool to optimize the fitting of a hearing aid,” J. Speech Hear. Res. **38**, 726–736.
Preminger, J. E., Neuman, A. C., Bakke, M. H., Walters, D., and Levitt, H. (**2000**). “An examination of the practicality of the simplex procedure,” Ear Hear. **21**, 177–193.
Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). “Speech recognition with primarily temporal cues,” Science **270**, 303–304.
Stelmachowicz, P. G., Lewis, D. E., and Carney, E. (**1994**). “Preferred hearing aid frequency responses in simulated listening environments,” J. Speech Hear. Res. **37**, 712–719.
Takayagi, H. (**2001**). “Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation,” Proc. IEEE **89**, 1275–1296.
Wakefield, G. H., van den Honert, C., Parkinson, W., and Lineaweaver, S. (**2005**). “Genetic algorithms for adaptive psychophysical procedures: Recipient-directed design of speech-processor MAPs,” Ear Hear. **26**, 57S–72S.