Optimization Models of Sound Systems Using Genetic Algorithms

Jinyun Ke*
City University of Hong Kong

Mieko Ogura†
Tsurumi University
University of California at Berkeley

William S-Y Wang‡
City University of Hong Kong
University of California at Berkeley

In this study, optimization models using Genetic Algorithms are proposed to study the configuration of vowels and tone systems. Similar to previous explanatory models that have been used to study vowel systems, certain criteria, which are assumed to be the principles governing the structure of sound systems, are used to predict optimal vowels and tone systems. In most of the earlier studies only one criterion has been considered. When two criteria are considered, they are often combined into one scalar function. The GA model proposed for the study of tone systems uses a Pareto-ranking method which is highly applicable for dealing with optimization problems having multiple criteria. For optimization of tone systems, perceptual contrast and markedness complexity are considered simultaneously. Although the consistency between the predicted systems and the observed systems is not as significant as those obtained for vowel systems, further investigation along this line is promising.

1 Introduction

Studies of the universal characteristics of sound systems in human languages can be pursued from two different approaches, inductive and deductive. The inductive approach is to analyze the database built from a survey of a large number of languages to arrive at a list of "universal" features which can be widely observed in the database. The deductive approach hypothesizes a number of principles related to speech production and perception processes, and predicts possible systems using these principles. These two approaches, however, are often interwoven. The principles hypothesized by the deductive approach

* Department of Electronic Engineering, City University of Hong Kong, Hong Kong
† Linguistics Laboratory, Tsurumi University, Yokohama, Japan; Project on Linguistic Analysis, University of California at Berkeley
‡ Department of Electronic Engineering, City University of Hong Kong, Hong Kong; Project on Linguistic Analysis, University of California at Berkeley

are modified or falsified by comparing the predictions with the results from the inductive analysis of real language systems. At the same time, the ultimate aim for inductive analysis is to seek intrinsic mechanisms and principles of human speech to explain the universals found in real systems.

For the inductive approach in phonological studies, there are two large scale databases available. One is the Stanford Phonology Archiving (SPA) Project (Vihman, 1977), which initially included 196 languages and was extended to 209 languages in 1978. The other is the UCLA Phonological Segment Inventory Database (UPSID) (Maddieson, 1984), which initially included 371 languages and was later extended to 451 languages (Maddieson and Precoda, 1990; Ladefoged and Maddieson, 1996). Many typological studies have been carried out based on these two databases. For example, in studying vowel systems, Crothers (1978) reported an analysis using the SPA database. Ladefoged and Maddieson (1990) and Schwartz et al. (1997b) reported comprehensive analyses for the vowels systems in UPSID.

Along with typological studies of the languages in these databases, explanatory models, which attempt to explore the intrinsic reasons for structures and universals, have also been proposed. In the study of vowel systems, the principle of *maximal perceptual contrast* has a long tradition in linguistics (Jakobson, 1941; Wang, 1968). The principle suggests that the vowel system tends to achieve a maximum contrast among the vowels in the system. A number of numerical studies adopting this principle have been proposed (Liljencrants and Lindblom, 1972; Crothers, 1978; Lindblom, 1986). Lindblom (1986) proposed the *sufficient perceptual contrast* principle under which more systems are predicted to be consistent with natural systems than those predicted by the maximal perceptual contrast principle. Boë, Schwartz, and Vallée (1994) and Schwartz et al. (1997a) added a new consideration called the *focalization* principle which is based on the observation that vowels with strong formant convergence would be perceptually preferred. More recently, de Boer (1997; 2000; 2001) proposed a synthesized model in which agents interact with each other through iterative imitation games. With explicit optimization, agents can develop coherent vowel systems which are close to real systems.

The works cited above are all concerned with vowel systems. Other components of a sound system, including consonants, tones (in tone languages), and pitch accent (in non-tone languages), have far fewer studies reported than those on vowels. Lindblom and Maddieson (1988) reported a study on phonetic universals in consonant systems using data from UPSID. They proposed that the structure of consonant systems does not arise from a single principle such as the maximization of perceptual contrast. Instead, articulatory factors interact with perceptual factors. According to their proposal, consonant inventories tend to evolve so as to achieve maximal perceptual distinctiveness at minimum articulatory cost.

There are some inductive studies on the universals of tone systems as well. For example, Maddieson (1978) reviewed the phonological universals of tones by analyzing data from SPA. Also, Cheng (1973) reported a detailed analysis of the tone systems in Chinese dialects. However, we have not found any explanatory models using a deductive approach for tone systems as those performed for vowel systems.

More recently, Redford, Chen, and Miikkulainen (2001) reported their studies on the universal and variations of syllable structures, i.e. the combinations of vowels and consonants. They developed a computational model, which is based on a version of the Genetic Algorithm (GA) (Holland, 1975), to simulate the emergence of syllable systems. A set of functional constraints related to perceptual distinctiveness and articulatory ease are taken into account as optimization objectives.

In this study, we report some optimization models using Genetic Algorithms to study optimal vowel and tone systems. In these models, the optimal systems are derived from the models based on various explicit optimization criteria, and compared with observed systems. In the study of vowel systems, we compare two sets of criteria, one considering only the principle of maximal perceptual contrast (Liljencrants and Lindblom, 1972), and the other considering both inter-vowel's perceptual distance and intra-vowel's spectral salience, that is, the dispersion-focalization principle proposed by Schwartz et al. (1997a). In the second set, the two objectives are combined into a scalar function. Comparing our results with earlier studies, the GA models demonstrate the effectiveness of the GA method in identifying the optimal systems based on the above criteria.

Secondly we apply the GA method to study tone systems. Two objectives, i.e. maximum perceptual contrast and minimum markedness complexity, are taken into account to predict the "optimal" tone systems. Instead of combining the two objectives into one fitness function, we use a Multi-Objective Genetic Algorithm model in which a Pareto-ranking method is applied for the fitness function. In order to make a comparison, we also try a simple GA model which uses perceptual distance only as the optimization criterion. The predicted systems are compared with the real systems for the two sets of criteria.

In the following parts of the paper, Section 2 gives a brief introduction to a simple Genetic Algorithm and a Multi-Objective GA. Section 3 reports the simulation for vowel systems and comparisons with previous reports. Section 4 introduces the models for tone systems, together with a new analysis of an available tone systems database. Conclusions and discussion are given in Section 5.

2 Introduction to Genetic Algorithms

2.1 Simple Genetic Algorithm

Genetic Algorithms (GAs) were first proposed by John Holland in the 1960s (Holland, 1975) and have become widely used in various disciplines. The original goal of Holland's GAs was to formally study the phenomena of adaptation by importing the mechanisms of natural adaptation into computer simulation models. However, most of the current applications of GAs are used for specific optimization problems, where the focus is on the derivation of optimal solutions to the problem rather than the process of adaptation.

The basic idea of GAs is based on "natural selection", the principle of "survival of the fittest", which assumes that the individual which is fitter to the environment produces more offspring; its "fit" genes are then transmitted to the next generation. A Genetic Algorithm operates on a population of chromosomes, each generating a potential solution to the studied problem. The process of a traditional simple GA is as follows: at the beginning, a population is randomly initialized, and the fitness of each chromosome is evaluated according to an objective, also called fitness, function. A number of chromosomes are selected as parents from the population according to their fitness, and parents then undergo crossover and mutation to produce offspring with certain probabilities. Offspring with better fitness are inserted into the population, replacing the inferior chromosomes in the last generation. With this replacement, usually the population size is kept constant. This cycle is repeated for a given number of generations, or stopped when a solution is obtained as optimal. This process leads to the evolution of a population in which the individuals are more and more suited to their environment, just as in natural adaptation. Due to its global search mechanism, a GA model usually can find the global optimal solutions in a more efficient way than traditional optimization methods.

2.2 Multi-objective Genetic Algorithm

In a traditional GA, the fitness function deals only with one optimization objective. However, many practical problems are concerned with several equally important, and usually conflicting, objectives. These types of problems are called Multi-objective or Multi-criteria Optimization Problems (MOP) (Stadler, 1988).

Human language is such a case of an MOP. A language system is constrained by many demands and requirements. We can consider that the current language system is the product of an optimization process based on such constraints. The constraints can be divided mainly into three categories: the speaker constraint, the listener constraint and the learner constraint, which often lead to different directions of development of the system. For example, for a sound system, the requirement from speaking and listening often conflict with each other. A sound which is easy for the speaker to produce may not be easy for the listener to perceive. Similarly, perceptually distinctive sounds may be difficult to pronounce. A system with a high perceptual contrast may have a high production cost at the same time, such as the consonant set [ɗ k’ ts ɬ m r ɟ] suggested by J. Ohala in questioning the effectiveness of the principle of maximum perceptual difference in explaining consonant universals (Lindblom and Maddieson, 1988).

We can see the effects of such a tug-of-war in various aspects of a language system. For example, in the perception of tones, a completely level tone is the easiest to differentiate from non-level tones from a psychophysical viewpoint. However, it requires much effort of the speaker to produce a perfectly level tone. As a consequence of accommodating a speaker’s effort, the listeners will shift their linguistic perception boundary between level and rising tones away from the psychophysical boundary, to allow some freedom in the articulation of the speaker (Wang, 1976).

Also, in syntax, a language with free word order may give the speaker a high flexibility in constructing sentences; however, it would place the burden on the listener to figure out the relationships among the words. This is solved by signaling the roles of words by various case markers. However, if the case marking system is too complex, it will be hard for the children to learn. Therefore there may exist a balance point among the different constraints of the three parties involved.

The most distinctive characteristic of an MOP is that the problem does not have one singular optimal solution, but rather a set of non-dominated[1], alternative solutions, which is often called the Pareto-optimal set. Recently a set of algorithms, called Multi-Objective Genetic Algorithms (MOGAs), have been developed specifically to solve such multi-objective problems. MOGAs have received much attention and many scientific and engineering applications have been reported (Fonseca and Fleming, 1998; Van Veldhuizen and Lamont, 2000).

The simplest and most common way to tackle an MOP is to combine the several objectives into one scalar function as the fitness function. Different objectives are given different weights based on some *a priori* knowledge (Stadler, 1988). However, such knowledge is not available very often, and most of the time the weights are chosen by trial and error. Thus the performance of the algorithm usually is sensitive to or biased by the weights. Early studies on sound system optimization with multiple criteria, such as Redford, Chen, and Miikkulainen (2001) and Schwartz et al. (1997a), adopted this approach.

Within the GA approach, another method called Pareto-ranking is often used in the fitness evaluation. The several objective values of a chromosome are kept as a vector, instead of being combined by a scalar function into one single fitness value. The fitness of a chromosome is determined by its ranking in the population which is obtained from comparing the its objective vector with others. There are many ranking methods and in this study we use the one proposed in Goldberg (1989). This method assigns ranks according to the following procedures. First, find the non-dominated chromosomes from the whole population, assign rank 1 to them, and remove them from further consideration in the ranking process afterwards. Then, find another set of non-dominated chromosomes

1 Assuming a minimization problem with p objectives, dominance is defined as follows: x_1 is said to dominate x_2 (or x_2 is inferior to x_1), if the fitness of x_1, $f(x_1)$, is partially less than the fitness of x_2, $f(x_2)$, i.e. $f_i(x_1) \leq f_i(x_2)$, $\forall i \in \{1, 2, \ldots, p\}$; and $f_i(x_1) < f_i(x_2)$, $\exists i \in \{1, 2, \ldots, p\}$. A non-dominated solution is such a solution that there are no other solutions whose objectives are all better than its.

from the remaining population and assign them rank 2, and so forth. To illustrate the algorithm, Figure 1 gives an example. Points A, B, C, D, E, F, G represent candidate solutions to a problem whose goal is to minimize two objectives. The objective values of the solutions are shown in the f_1-f_2 plane. According to the method, solutions A, B, C and D are all non-dominated and therefore assigned rank 1. Solutions E and F both have rank 2, while G has 3, the worst rank.

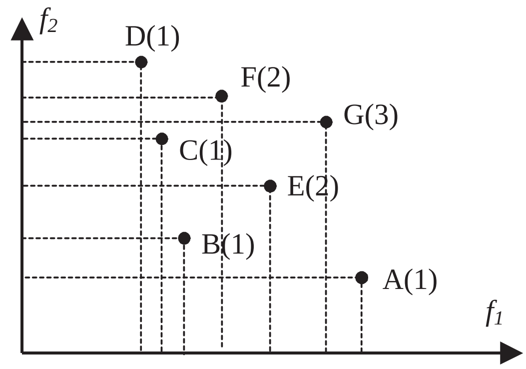

Figure 1
An illustration of Goldberg's Pareto ranking method. The numbers in parentheses represent the ranks for the chromosomes.

3 Optimization model for vowel systems

In this study, we use a simple GA model to search for the optimal configuration of systems of simple vowels. We apply various optimization criteria proposed by earlier studies (Liljencrants and Lindblom, 1972; Lindblom, 1986; Schwartz et al., 1997a), in our GA model and compare the predictions. Also different vowel inventories are used to provide another set of comparison. The predictions are also compared with observed systems.

3.1 Implementation of the GA model

The GA model consists of a population with a number of chromosomes, each representing a possible vowel system. Each vowel is encoded by the three primary articulatory parameters, that is, tongue height, tongue backness, and lip roundness. The first two articulatory parameters are supposed to be continuous within the range of [0,1], while the last parameter is a binary value. Though this encoding method allows an infinite number of vowels, following previous studies (Lindblom, 1986; Schwartz et al., 1997a), we assume there is a limited inventory of prototypes from which the system can select candidate vowels. Only normal plain vowels are considered. The encoding of the prototypical vowels is designed according to the vowel position shown in the International Phonetic Alphabet(IPA) vowel chart. Although the IPA vowel chart is better interpreted as an acoustic chart, rather than an accurate projection of real articulation of the vowels, the chart can still be assumed to reflect the relative positions for articulation.

Two inventories of vowel prototypes are used, denoted as INV_L and INV_S respectively. One consists of 18 vowels from a set of 19 vowels given in Lindblom (1986)[2]. The other set includes 24 vowels extracted from the set of 33 vowels given in Schwartz et al. (1997a). Figure 2 shows the two inventories in terms of the vowels' first formants (F_1) and transformed second formants (F_2') (Fant, 1966) (see the Section 3.2 for more explanation), both expressed in terms of the Bark scale (Hartmann, 1997).

In the GA model, one-point crossover and one-point mutation are used. Take the simulation of 3-vowel systems as an example. Two chromosomes are selected as parents from the population, as shown in Figure 3. Parent 1 includes three vowels: ʉ ø and ɛ, represented by [0.0 0.5 1], [0.3 0.1 1], and [0.7 0.2 0] respectively. And three vowels ɯ, ə, and ɑ are included in parent 2, represented by [0.0 1.0 0 | 0.5 0.5 0 | 1.0 1.0 0]. Crossover is randomly chosen to take place between two vowels, say the second and the third vowel in the example, and the two chromosomes exchange their vowels. Next, by random, the mutation occurs to the third vowel in the second offspring, [ɛ] is changed to [œ]. So the

2 The original inrounding front vowel [y] is deleted as it is not a common primary vowel, and the outrounding [ü] is changed to symbol [y] in order to conform to the IPA transcription.

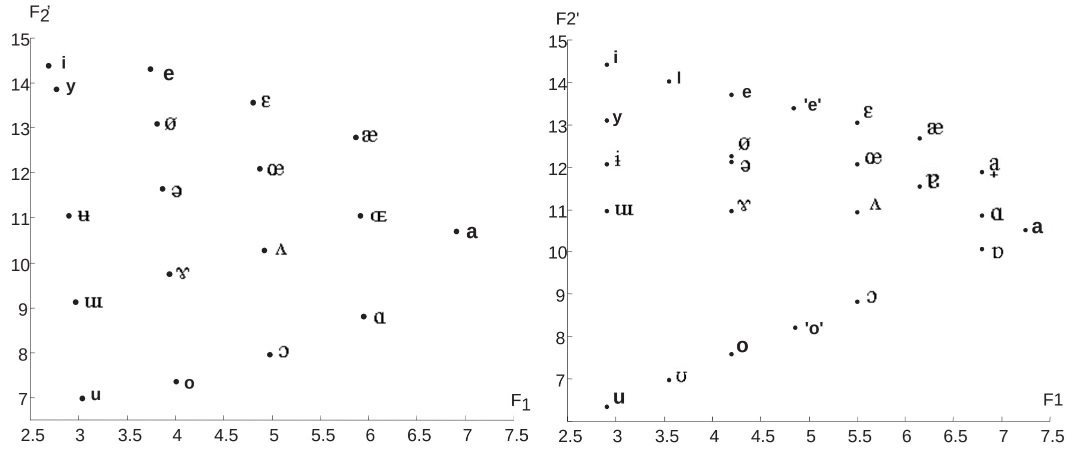

Figure 2
The F_1 and F_2' diagram of prototypical vowels, left: 18 vowels from INV_L; right: 24 vowels from INV_S.

two offspring generated from the pair of parents are [ʉ, ø, ɑ] and [ɯ, ə, œ].

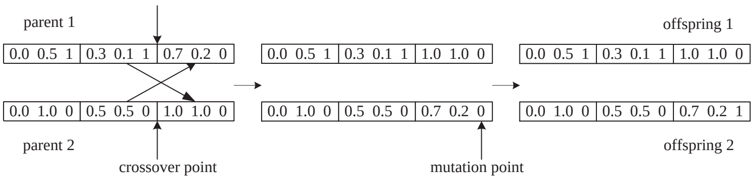

Figure 3
Crossover and mutation operations in simulation of 3-vowel systems.

Since the aim of the GA model used here is to find the optimal solution, the crossover and mutation rates are both set to 1.0 in order to have the highest efficiency in searching for the optimal solution. If the genetic operations generate offspring with one vowel occurring twice in a system, this offspring is removed from the population and a new chromosome is randomly generated to keep the population size constant. Next, offspring with higher fitness are re-inserted into the population, replacing those individuals with lower fitness values.

3.2 Fitness evaluation functions

Two sets of criteria are taken into account. One considers only the principle of maximal perceptual contrast (Liljencrants and Lindblom, 1972), and the other considers both inter-vowel's perceptual distance and intra-vowel's spectral salience related to the proximity of formants, i.e. the dispersion-focalization principle proposed by Schwartz et al. (1997a).

For the first criteria, the objective is to minimize the following fitness function:

$$\mathcal{F}_1 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{d_{ij}^2} \tag{1}$$

where d_{ij} is the perceptual distance between vowels i and j. Various metrics for calculating perceptual distance between vowels have been proposed based on perceptual experiments manipulating different combinations of formants and amplitudes of speech signals (Schwartz et al., 1997a). Usually the acoustic parameters, i.e. the higher formants (F_2, F_3, and F_4), are first combined and transformed into an "equivalent second formant" F_2'. And the auditory distance between two vowels is calculated as the weighted Euclidean distance in the space of F_1 and F_2', where the weight between F_1 and F_2' is determined by λ:

$$d_{ij} = \sqrt{(F_{1i} - F_{1j})^2 + \lambda(F'_{2i} - F'_{2j})^2} \tag{2}$$

In this study, two methods of calculating F_2' are tried, one proposed in Fant (1966), and one given in Schwartz et al. (1997a), in order to examine the effect of different metrics of calculating perceptual contrast. Therefore, we will have two fitness functions regarding the first criterion perceptual contrast, denoted as $\mathcal{F}_{1F}$ and $\mathcal{F}_{1S}$.

The second criteria includes another objective in addition to the above $\mathcal{F}_1$, which is the intra-vowel formant convergence:

$$\mathcal{F}_c = \sum_{i=1}^{n} \frac{-1}{(F_{2i} - F_{1i})^2} + \sum_{i=1}^{n} \frac{-1}{(F_{3i} - F_{2i})^2} + \sum_{i=1}^{n} \frac{-1}{(F_{4i} - F_{3i})^2} \tag{3}$$

The overall fitness function for the second set of criteria is a weighted summation of the above two objectives, i.e.

$$\mathcal{F}_2 = \mathcal{F}_1 + \alpha \mathcal{F}_c \tag{4}$$

The values of λ and α are crucial for the prediction. Schwartz et al. (1997a) tested a number of values and found that the following ranges give the best prediction with their vowel inventory[3]: $0.04 \leq \lambda \leq 0.09$ and $0 \leq \alpha \leq 0.4$. In our experiments, we choose the values $\lambda = 0.0625$ and $\alpha = 0.3$ which are within the above ranges.

3.3 Results and analysis

We predict the optimal 3- to 7-vowel systems using six sets of experiments, each for a combination of one of the two vowel inventories (INV_1 and INV_2), and one of the three different fitness functions ($\mathcal{F}_{1F}$, $\mathcal{F}_{1S}$ and $\mathcal{F}_2$). The predicted systems are listed in Table 1, together with the commonly observed systems found in the database UPSID given in Schwartz et al. (1997a), and the predictions given in Schwartz et al. (1997a) using the same parameters as those for $\mathcal{F}_2$ here, listed as S_0.

First we compare the predictions using the same vowel inventory but different fitness functions. The two perceptual distance metrics ($\mathcal{F}_{1F}$ and $\mathcal{F}_{1S}$) produce the same predictions for systems of small sizes, i.e. 3-, 4- and 5-vowel systems, but different predictions for larger systems, i.e. 6- and 7-vowel systems, which means that predictions for larger systems are more sensitive to the transformations $F2'$. That transformations used in Schwartz et al. (1997a) produces a more spread perceptual space in the $F2'$ dimension, and especially [i] has a much larger $F2'$ than Fant's transformation. It is hard to give an overall evaluation of which perceptual distance metric gives better predictions based on these results. $\mathcal{F}_{1S}$ predicts a symmetric 6-vowel system while $\mathcal{F}_{1F}$ does not when using

3 Note that the λ in our formula 2 corresponds to λ^2 in Schwartz et al. (1997a), therefore the range of λ is modified accordingly.

Table 1
Frequent vowel systems in UPSID and predicted vowel systems using different inventories and fitness functions.

N	observed systems	predicted systems	
		INV_S	INV_L
3	[i, a, u](14)	$\mathcal{F}_{1F}$:[i, a, u] $\mathcal{F}_{1S}$:[i, a, u] $\mathcal{F}_2$: [i, a, u] S_0: [i, a, u]	[i, a, u] [i, a, u] [i, a, u]
4	[i, ‘e’, a, u](14) [i, a, u, ɨ](5) [i, a, ‘o’, u](2) [e, a, o, ə](2)	$\mathcal{F}_{1F}$:[i, ɛ, a, u] $\mathcal{F}_{1S}$:[i, ɛ, a, u] $\mathcal{F}_2$: [i, ‘e’, a, u] S_0: [i, ‘e’, a, u]	[i, ɛ, a, u] [i, ɛ, a, u] [i, ɛ, a, u]
5	[i, ‘e’, a, ‘o’, u](97) [i, ɛ, a, u, ɨ](3)	$\mathcal{F}_{1F}$:[i, ɛ, a, ‘o’, u] $\mathcal{F}_{1S}$:[i, ɛ, a, ‘o’, u] $\mathcal{F}_2$: [i, ɛ, a, ‘o’, u] S_0: [i, ‘e’, a, ‘o’, u]	[i, ɛ, a, ɔ, u] [i, ɛ, a, ɔ, u] [i, ɛ, a, ɔ, u]
6	[i, ‘e’, a, ‘o’, u, ə](26) [i, ‘e’, a, ‘o’, u, ɨ](12) [i, ‘e’, æ, ɑ, ‘o’, u](12) [i, e, a, ɔ, o, u](4)	$\mathcal{F}_{1F}$:[i, e, æ, a, ‘o’, u] $\mathcal{F}_{1S}$:[i, e, æ, a, o, u] $\mathcal{F}_2$: [i, e, æ, a, ‘o’, u] S_0: [i, ɛ, a, ‘o’, u, ɨ]	[i, ɛ, a, ɔ, u, ʉ] [i, æ, a, ɔ, u, ø] [i, ɛ, a, ɔ, u, ʉ]
7	[i, e, ɛ, a, ɔ, o, u](23) [i, ‘e’, æ, a, ‘o’, u, ə](6) [i, ‘e’, a, ‘o’, u, ə, y](5) [i, ‘e’, a, ‘o’, u, ɨ, ə](4) [i, e, ɛ, a, ‘o’, u, ɨ](3)	$\mathcal{F}_{1F}$:[i, ‘e’, æ, a, ‘o’, u, ɯ] $\mathcal{F}_{1S}$:[i, e, æ, a, ɔ, u, ɯ] $\mathcal{F}_2$: [i, e, æ, a, ‘o’, u, ɯ] S_0: [i, e, ɛ, a, ‘o’, u, ɯ]	[i, e, ɛ, æ, a, ɔ, u] [i, e, æ, a, ɔ, u, ʉ] [i, e, æ, a, ɔ, u, ʉ]

INV_S; however, when using INV_L, $\mathcal{F}_{1S}$ predicts a strange 6-vowel system with a front rounded vowel ø which is rarely attested in primary vowel systems while $\mathcal{F}_{1F}$ predicts a system close to the observed system.

From the comparison of the two fitness functions $\mathcal{F}_{1S}$ and $\mathcal{F}_2$, we can see that the predictions in most cases are the same. There are some exceptions, i.e. in those experiments using INV_S for 4-, 6- and 7- vowel systems, and INV_L in 6-vowel systems. Predictions given by $\mathcal{F}_2$ only differ in some small variations between ɛ and ‘e’, ‘o’ and o, ɔ and ‘o’. $\mathcal{F}_2$ does not predict more non-peripheral vowels than $\mathcal{F}_{1S}$, inconsistent with the proposal of Schwartz et al. (1997a), although the parameters are set within the optimal range. The discrepancies may need further examination, because in our experiment using the same $\mathcal{F}_2$ and INV_S as those in Schwartz et al. (1997a), the predictions of optimal

5-, 6- and 7-vowel systems are different from their reports (reproduced as the list of $S0$ in the table), given $\lambda = 0.0625$, $\alpha = 0.3$. However, when λ is set to 0.09 for the 5-vowel system, and $\lambda = 0.025$ and $\alpha = 0.1$ for the 6-vowel system, the predictions are the same as those in Schwartz et al. (1997a).

Second, we compare predictions with the same fitness function but different inventories. Though the original dataset where INV_S is selected from is said to be carefully controlled in order to sample the acoustic space as evenly as possible (Schwartz et al., 1997a), we do not see much difference between INV_S and INV_L for predicting small size systems. If we consider the vowel ɛ in Lindblom's inventory to be equivalent to the vowel 'e' in Schwartz et al. (1997a), ɔ to 'o', and ʉ to ɯ, then the predictions from the two inventories are almost the same. This is not surprising since we can see from Figure 2 that the peripheral vowels in the two vowel inventories are almost the same in the F_1-$F2'$ plane. However, for the 6-vowel system with three fitness functions and the 7-vowel systems with the first fitness function ($\mathcal{F}_{1F}$), there are some big differences. This may be mainly due to the fact that the non-low unrounded back vowels are much farther away from rounded back vowels in INV_S than in INV_L.

Compared the predicted systems with the observed systems, we find that only the most frequently observed 3- and 4-vowel systems are predicted, while other predictions do not match the observed systems. The reasons may include the following: first, the vowel inventories, specially INV_L, do not provide enough vowel prototypes, such as 'e' and 'o' which are not included in INV_L; second, the perceptual distance metric used in the study may have not perfectly reflected the actual human perception mechanism; third, only one optimal system can be identified using the current simple GA model. Further considerations of the optimization criteria for the GA model including the *sufficient*, instead of *maximal*, perceptual contrast principle proposed by Lindblom (1986) may lead to more consistent predictions. Also the central vowel ə, which occurs often in large vowel systems, may call for another optimization criterion (Schwartz et al., 1997a). GA models are promising in carrying out such investigations in which multiple optimization criteria are addressed simultaneously. The following study on the optimal tone systems is such an experiment.

4 Optimization model for tone systems

A tone language is a language having lexically contrastive pitch on each syllable (Pike, 1948). Tone languages are found in many parts of the world (Wang, 1991). Though tone as fundamental a constituent in tone languages as are vowels and consonants, to our knowledge, there has been no explanatory or numerical model proposed to study the universal structure of tone systems such as those being done for vowel and consonant systems. In this section, we extend the GA models reported above for vowel systems to study the configuration of tone systems from the optimization perspective. Two different sets of criteria are investigated. The first criterion considers only the perceptual contrast, as was done for the vowel systems; the second takes both perceptual contrast and markedness complexity into account. The predicted systems are analyzed and compared with those reported from empirical studies.

4.1 Tone inventory and chromosome representation

Similar to the simulation of vowels, we first choose a tone inventory from which a system selects individual tones. Wang (1967) suggested 13 idealized tones in his study of the phonological features of tones, including five level (11 22 33 44 55), two rising (35 13), two falling (53 31), two falling-rising (535 313), and two rising-falling (353 131)[4]. These 13 tones are considered to represent the maximum contrasts found in any language. Later in

4 Tones are represented according to the conventional Chao's five-level transcription system (Chao, 1930).

a tone perception experiment, Gandour (1983) used an extended set of 19 tones, adding to Wang's list two rising tones 15 and 24, two falling tones 51 and 42, and two complex tones 424 and 242. The 19 tones are shown in Figure 4. In this study, we take Gandour's 19 tones as the inventory.

In the model, the chromosome is represented by a number of tones each of which is selected from the 19 tones. Each tone is described by three numerals representing its shape. The genetic operations and parameters are the same as those in the vowel models.

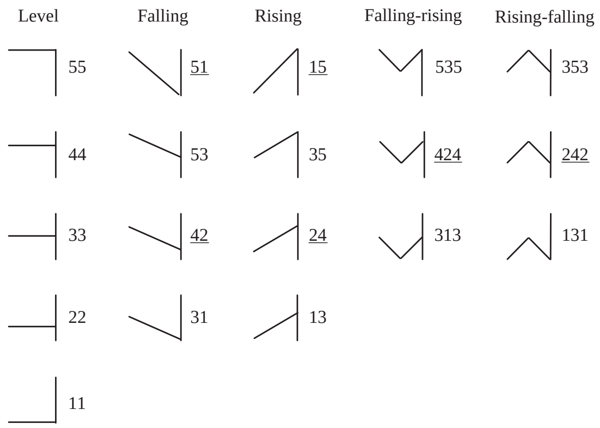

Figure 4
Numeric and corresponding graphic representation of the 19 tones proposed in Gandour (1983). Tones not found in Wang (1967) are underlined.

4.2 Fitness evaluation functions

4.2.1 The first objective It is assumed that tone systems tend to have a maximum perceptual contrast within the system, much like that proposed for vowel systems. Similar to the study of vowel systems, we need to derive a method to calculate the perceptual distance between tones. In this study, we use the experimental results from Gandour (1983) to develop such a metric to measure the perceptual distance.

Gandour's experiment was designed to investigate the perceptual dimension of tone and the effect of linguistic experience on a listener's perception of tone. He synthesized speech-like monosyllables [wa] with the 19 types of tones superimposed on them. Four groups of subjects who were from four tone languages including Cantonese, Mandarin, Taiwanese and Thai, and one group from a non-tone language, English, made judgments of dissimilarity between paired stimulus tones. The collected data were analyzed by an INDSCAL (Individual Differences SCALing) model, and the perceptual dissimilarity for the 19 tone types is presented in a perceptual space shown in Figure 5. Using this result, we design a metric of calculating the perceptual distance of a pair of tones i and j as computing their Euclidean distance in the perceptual plane:

$$d_{ij} = \sqrt{(Dim_{1i} - Dim_{1j})^2 + (Dim_{2i} - Dim_{2j})^2} \tag{5}$$

Dim_{1i} and Dim_{2i} represent the two coordinates of tone i in the perceptual plane. Similar to the method used in the vowel systems, the perceptual contrast within a tone system is measured by the total perceptual distance for all pairs of tones. The fitness function is therefore the same as the $\mathcal{F}_1$ given in Section 3.2.

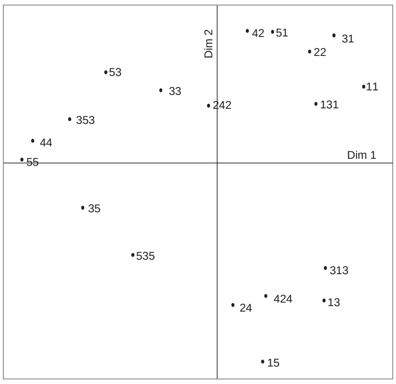

Figure 5
Dimensions 1 and 2 of the two-dimensional INDSCAL tone space (Adapted from Gandour (1983))

4.2.2 The second objective Following the first objective of considering perceptual contrast, the second consideration would naturally be the production cost. Various mechanisms for controlling the tension of vocal cords and subglottal air pressure and their effect in regulating pitch change have been proposed (Ohala, 1978). Different laryngeal muscles, such as the cricothyroid, sternohyoid and sternothyroid muscles, etc. are found to perform various actions in raising or lowering pitch. It is found that there is an asymmetry in the maximum speed of rises as compared to falls in pitch change, which suggests that falling tone may in some sense be easier to produce than rising tones (Collier, 1984; Ohala, 1978). It is found that level tones are universally preferred to contour tones, and simple contour tones to complex contour tones (Maddieson, 1978), which may be due to the different production cost of various types of tones. However, it is hard to quantify these differences, and no systematic measurements have been available yet.

Due to the lack of data measuring the production effort of different tones, we choose another criterion as the second objective: the markedness complexity, which is based on a study of phonological features of tones proposed in Wang (1967). Each tone is assigned a complexity value based on the analysis of the tones with features and marking conventions, as shown in Table 2. The 'm' stands for the marked specification, which is the favored specification, while 'u' for unmarked. The assignment of '+' and '-' is due to the consideration that there is no empirical ground for favoring either +HIGH or -HIGH, or RISING or FALLING. We note that the latter may need further justification or modification regarding our earlier discussion above on the rising and falling tones. In this study, however, we still adopt this analysis for our simulation. The specifications 'm', '+' and '-' each add one unit to the complexity, while 'u' does not.

Table 2
Relative complexity of tones as defined by marking conventions(adapted from Wang (1967)).

	55	11	44	22	33	35 /15	13 /24	53 /51	31 /42	535	313 /424	353	131 /242
CONTOUR	u	u	u	u	u	m	m	m	m	m	m	m	m
HIGH	+	-	+	-	-	+	-	+	-	+	-	+	-
CENTRAL	u	u	m	m	m	u	u	u	u	u	u	u	u
MID	u	u	u	u	m	u	u	u	u	u	u	u	u
RISING	u	u	u	u	u	+	+	-	-	+	+	+	+
FALLING	u	u	u	u	u	-	-	+	+	+	+	+	+
CONVEX	u	u	u	u	u	u	u	u	u	u	u	m	m
COMPLEXITY	1	1	2	2	3	4	4	4	4	4	4	5	5

While maintaining the original complexity assignment of the 13 typical tones given in Wang (1967), we have incorporated the six additional tones proposed by Gandour (1983). It is assumed that tone pairs such as 35 and 15, 13 and 24, 53 and 51, 31 and 42, 313 and 424, 131 and 242, are of the same complexity. As far as we are aware, there is no tone system having more than two falling or rising contrasts which all give lexical distinctions. Those transcriptions of tone systems which do incorporate more than two rising or two falling tones may be due to over-differentiation within a single tone paradigm (Wang, 1967).

Markedness is a method of representing the linguist's knowledge of a phonological system. This knowledge derives primarily from observations of three sorts: the frequency of distribution of the sounds in the languages of the world; the patterns of historical change in sound systems; and the acquisition of sounds in children and the dissolution of sounds in linguistic pathology. Therefore the assigned complexity of tones based on the markedness may reflect an integrated effect of perception, production and learnability.

The two objectives, perceptual distance and markedness complexity, are taken into account in a Multi-objective GA model using the Pareto-ranking method as introduced in Section 2.2, to predict optimal tone systems. Simulation results are shown in the section following the empirical data analysis.

4.3 Empirical data analysis

Before reporting the simulation results, we report our analysis on an available database of observed tone systems with which we can compare our predictions. To our knowledge, there is no large database for tone systems of the same scale as SPA or UPSID for vowel and consonant systems. However, we have been able to find a a computer database consisting of 737 dialect[5] locations in China, which was compiled by C. C. Cheng (1973). Some of the 737 entries are from the same location, but from different reports or at different times. They are considered as individual systems in our analysis since they were obtained independently and we want to consider as many systems as possible in our analysis, though it is possible that our analysis may be contaminated by the unequal quality of the reports.

In order to make comparison with predicted systems, we normalize non-typical tones in Cheng's database to the 19 typical ones. For example, tones 54, 12, and 324 are converted to 53, 13, and 313 respectively. Such normalization also provides some advantage in dealing with different types of transcription of the same system. Very often different researchers each have their own individual strategy of transcription, and different informants employed for the same dialect also have individual differences. Therefore there often arise controversies such as in Cantonese whether a tone is 13 or 23. By normalizing the tones into the 19 typical tones, some of such controversies will be resolved. Table 3 shows the frequencies of occurrence of the 19 types of tone in the normalized database. After normalization, when the changes result in a situation such that one tone occurs twice in one system, these systems were excluded in the normalized database for further analysis. The original database includes 737 systems of 606 different types. After normalization, there are 641 systems of 319 different types. Table 4 shows the frequencies of different sizes of tone system in both the original and the normalized database. Systems with 4 tones are by far the most frequent type.

Figure 6 shows the frequencies of the tone types occurring in the 4-tone normalized systems. Frequencies of occurrence are indicated beside the tone. We can see that tones are heavily clustered in three areas in the perceptual plane, that is, the upper right

5 The "dialects" we used here are referred to as the various languages such as Mandarin, Cantonese, Min dialect, Wu dialect, ect.

Table 3
Frequencies of occurrence of the 19 tone types in the normalized database.

tone	31	53	55	42	35	24	313	44	33	13
freq	326	294	291	281	262	248	241	219	185	167
tone	11	51	22	424	353	242	131	15	535	
freq	86	78	66	27	12	8	8	6	1	

Table 4
Frequencies of occurrence and types of different sizes of tone system in both original and the normalized database.

size	3	4	5	6	7	8	9	total
freq(original)	26	497	105	91	10	7	1	737
freq(normalized)	22	448	86	79	4	2		641
types(normalized)	19	162	72	60	4	2		319

(including tones 42, 51, 31, 22, 11, 131) –cluster 1; the low right (24, 13, 424, 313, 15) – cluster 2; and the mid left (55, 44, 35, 53, 33, 353, 242) – cluster 3. Tone 535, though not occurring in 4-tone systems, can also be classified into cluster 3 as inferred from Figure 5. We calculate the frequencies of occurrence of tones in these three areas for 3-, 4- and 5- tone systems, as shown in Table 5. It is found that most of the systems tend to select tones from each of the three individual clusters, rather than selecting tones from only one or two clusters. The percentages of systems having tones from each of the three clusters are shown in the last column of Table 5. We can infer from this observation that the observed systems tend to have large perceptual contrast within the systems, since the tones from distinctive clusters often have larger perceptual distance than tones in the same cluster. Furthermore, for 4- and 5- tone systems, the three clusters are not utilized evenly. It is obvious that cluster 2, which includes most of the rising tones (except high rising tone 35 which is contoured in cluster 3), contributes fewer tones than the other two clusters. This implies that rising tones are less preferred in the observed systems.

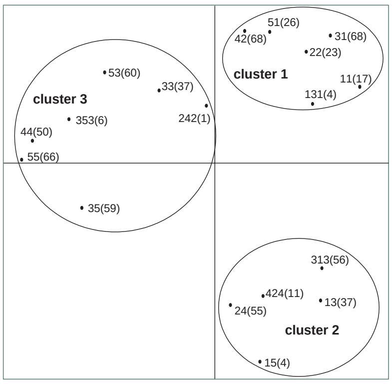

Figure 6
Frequency of occurring of tones in observed 4-tone systems shown in the perceptual space (total 162 systems)

Table 5
Number of tones in the three clusters in perceptual space and percentage of systems having tones from each of the three clusters.

	# of systems	tones in C1	tones in C2	tones in C3	% of systems having tones from each of the 3 clusters
3-tone system	19	19	17	21	73.7%
4-tone system	162	206	163	279	75.0%
5-tone system	72	107	91	162	74.3%

4.4 Simulation results and comparison with observed systems

Tables 6, 7 and 8 give the predicted "optimal" 3-, 4- and 5- tone systems obtained from the GA model. A number of frequent tone systems in the normalized database are also given in the tables for comparison. The frequencies of the tone systems observed in the normalized database are given in parenthesis besides the tones.

For a given size of system, the model using the single objective, i.e. maximal perceptual contrast, predicts only one optimal system which is indicated with asterisks in the table. This optimal system has the maximum perceptual contrast but the largest markedness complexity. On the other hand, the two-objective model predicts a number of optimal systems, besides the optimal one predicted by the single objective model. These optimal systems, as the points A, B, C, and D shown in Figure 1, form a Pareto-optimal set. They are equally good in terms of the two objectives, i.e. large perceptual contrast and small markedness complexity. The two-objective model may be viewed as a possible implementation of the "sufficient perceptual contrast" proposed for the study of optimal vowel systems by Lindblom (1986).

Table 6
Predicted optimal and frequent observed 3-tone systems.

predicted system3					observed system3				
perdist	complexity	t1	t2	t3	perdist	complexity	t1	t2	t3 (freq)
*14.41	9	55	31	15	12.97	9	55	31	424(3)
13.77	7	55	22	15	11.52	12	42	35	13(1)
13.97	6	55	11	15	8.84	11	33	53	313(2)

Table 7
Predicted optimal and frequent observed 4-tone systems.

predicted systems						observed systems					
perdist	complexity	t1	t2	t3	t4	perdist	complexity	t1	t2	t3	t4 (freq)
*24.63	12	55	11	31	15	22.42	14	44	53	31	13(21)
24.55	10	55	11	31	15	22.02	13	55	42	31	24(27)
23.67	9	44	22	11	15	21.98	13	55	53	31	24(17)
23.24	8	55	44	11	13	21.03	13	55	51	35	313 (20)

Table 8
Predicted optimal and frequent observed 5-tone systems.

predicted systems							observed systems						
perdist	compl	t1	t2	t3	t4	t5	perdist	compl	t1	t2	t3	t4	t5 (freq)
*39.20	17	55	53	31	15	13	35.90	13	55	44	22	31	13(1)
38.44	14	55	11	53	15	13	35.82	17	55	42	31	24	313 (4)
38.60	12	55	44	11	31	15	34.54	15	55	44	42	24	13(1)
38.17	10	55	44	22	11	15	32.33	12	55	22	11	31	24(1)
							32.15	10	55	44	33	11	24(1)
							28.12	17	11	51	53	42	24(2)
							27.09	18	44	53	42	31	35(3)

We can observe in the predicted systems similar characteristics as in the observed systems, i.e. the uneven utilization of the three clusters in the perceptual space. If we assume that tones within the same cluster are interchangeable, then we can see that most of the predicted optimal systems have correspondents in the observed systems, except one prediction for a 3-tone system (55,11, 31), which does not include have a tone from cluster 3.

The predicted systems seem to exploit more tones in the outer area of the perceptual space, such as 55, 11, 31 and 15, while in the observed systems tones in the relative inner area such as 35, 53, 42, 24 and 313 are more frequent. High rising tone 15 occurs often in predicted systems, which is due to its high salience in the perceptual space. In the observed systems, however, tone 15 is very infrequent (only 6 occurrences among the 641 systems). In the observed systems, pairs of contrasting tones, such as 24 and 42, 13 and 31, occur quite often. This tendency is not clear in our prediction. Tone 31 co-occurs more frequently with tone 15 than with tone 13, which again is due to the perceptual salience of tone 15.

The observed systems include more tones close to the center of the perceptual space, while the predicted systems prefer tones located in the periphery of the perceptual space. This is similar to the long-standing problem in the study of vowel systems (Liljencrants and Lindblom, 1972). In the observed vowel systems, especially the larger systems, central vowels commonly occur (examples can be seen in Table 1), while the proposed optimal vowel systems predict them only rarely. The utilization of less peripheral areas suggests that the role of maximizing perception contrast may need to be adjusted, or that more optimization criteria in addition to the perceptual contrast and markedness complexity should be added.

Moreover, the markedness complexity we hypothesize is an abstract measure which incorporates many factors, including perception and production. Thus the consideration of perception contrast may have been duplicated in the fitness function. This may be another reason for the discrepancy mentioned in the above paragraph. When more empirical studies in the physiology of tone production are available, we may consider tone production as an individual objective in the fitness function, which we would expect to allow better predictions.

5 Conclusions and discussion

In this study, we apply optimization models using Genetic Algorithms to study the configuration of vowels and tone systems. This approach is similar to previous explanatory models that have been used to study vowel systems. Certain criteria, which are assumed to be the principles governing the structure of sound systems, are used to predict the optimal systems. In most of the previous studies, only one criterion has been considered (Liljencrants and Lindblom, 1972; Crothers, 1978; Lindblom, 1986). When two criteria are considered, the two objectives are combined into a single weighted function (Boë, Schwartz, and Vallée, 1994). In our study of vowel systems, the simple GA model also adopts a weighted function to combine two criteria, perceptual contrast and focalization. In the study of tone systems, however, we apply a Multi-Objective GA model which uses Pareto-ranking method to consider two criteria, including perceptual contrast and markedness complexity, simultaneously, without combining them into a scalar function. Prior knowledge of the weights between the two criteria are not necessary.

Another advantage of an MOGA is that we can obtain a set of Pareto optimal results, instead of only one optimal. An MOGA model generates more optimal predictions than a single-objective model and therefore it is more likely to have more predicted systems close to the observed systems. Although the consistency between the current predictions and the observed systems is not as significant as that obtained for vowel systems, further investigation along this line is promising.

Following the deductive approach pursed in this study, we can design various criteria to predict optimal systems. This approach provides the convenience and freedom in the manipulation of different parameters in the models, such as the parameters λ and α in Schwartz et al. (1997a), to testify different hypothesized mechanisms. However, it is necessary to seek explanations for such parameters in terms of physiological or cognitive constraints.

Studies following the deductive approach must not be independent from the inductive approach. For example, in the study of tone systems, not many comprehensive tone databases are made available. The resources which our investigation of tone systems relies on, including the experiment in Gandour (1983) and the database in Cheng (1973), are mostly based on the observation of tone languages found in Asian languages. The incorporation of data from other types of tone languages in Africa and America is expected to help in refining our explanatory hypothesis about the configuration of the systems.

Lastly, we would like point out that though in this study we apply optimization to predict vowel and tone systems, we do not imply that there exists any explicit and/or global optimization processes in the formation of such systems. We have no grounds to believe that speakers are aware of what sounds will provide maximal perceptual contrast or require the least production effort, and therefore deliberately choose those sounds. The optimization must be an emergent property from the interactions of language users (de Boer, 2000; de Boer, 2001). Each individual speaker has certain physiological and cognitive constraints which limit the possible sounds and assign preference to certain sounds. However, these constraints only provide a range of possibilities. It is the interactions among individuals that determine precisely which systems will emerge. That is why different configurations, even sub-optimal ones in the sense of some hypothesized criteria, can be observed in real systems. Research including modeling from this perspective is promising and may lead to more realistic predictions.

Acknowledgments
This research is supported in part by two grants from the City University of Hong Kong, No. 7100096 and 9010001. The second author is also supported by a grant from the Ministry of Education, Science, Sports and Culture of Japan, No. 11610512. We thank Prof. C. C. Cheng for providing us with his tone database of Chinese dialects. We are thankful to Dr. Lisa Husmann and Dr. James Minett for their kind help in preparing this paper. Also we greatly appreciate the three reviewers for their very helpful comments and suggestions.

References

Boë, Louis-Jean, Jean-Luc Schwartz, and Nathalie Vallée. 1994. The prediction of vowel systems: Perceptual contrast and stability. In Eric Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*. Chichester: John Wiley & Sons, pages 185–213.

Chao, Yuan Ren. 1930. A system of tone letters. *Le Maître Phonétique*, 45:24–27.

Cheng, Chin Chuan. 1973. A quantitative study of chinese tones. *Journal of Chinese Linguistics*, 1(1):93–110.

Collier, Rene. 1984. Some physiological and perceptual constraints on tonal systems. In Bernard Comrie Brain Butterworth and Östen Dahl, editors, *Explanations for Language Universals*. Walter de Gruyter & Co, Berlin, pages 237–247.

Crothers, John. 1978. Typology and universals of vowel systems. In J. H. Greenberg, editor, *Universals of Human Language(Phonology)*, volume 2. Stanford University Press, Stanford, Calif., pages 93–152.

de Boer, Bart. 1997. Generating vowel systems in a population of agents. In Phil Husbands and Inman Harvey, editors, *Fourth European Conference on Artificial Life*, pages 503–510, Cambridge Mass. MIT Press.

de Boer, Bart. 2000. Self-organization in vowel systems. *Journal of Phonetics*, pages 441–465.

de Boer, Bart. 2001. *The Origins of Vowel Systems*. Oxford University Press, Oxford ; New York.

Fant, Gunnar. 1966. A note on vocal tract size factors and non-uniform f-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 1:22–30.

Fonseca, Carlos M. and Peter J. Fleming. 1998. Multiobjective optimization and multiple constraint handling with evolutionary algorithms—Part I: A unified formulation. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 28(1):26–37.

Gandour, Jack. 1983. Tone perception in far eastern languages. *Journal of Phonetics*, 11:149–175.

Goldberg, David E. 1989. *Genetic Algorithms in Search, Optimization, and Machine learning*. Addison-Wesley.

Hartmann, William M. 1997. *Signals, Sound, and Sensation*. AIP Press, New York.

Holland, John H. 1975. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.

Jakobson, Roman. 1941. *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala. Reprinted in Selected writings I, 328-401. The Hague: Mouton, 1962.

Ladefoged, Peter and Ian Maddieson. 1990. Vowels of the world's languages. *Journal of Phonetics*, 18:93–122.

Ladefoged, Peter and Ian Maddieson. 1996. *The sounds of the world's languages*. Blackwell, Oxford.

Liljencrants, Johan and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, 48:839–862.

Lindblom, Björn. 1986. Phonetic universals in vowel systems. In John J. Ohala and Jeri J. Jaeger, editors, *Experimental Phonology*. Academic Press, Inc., Orlando, Florida, pages 13–44.

Lindblom, Björn and Ian Maddieson. 1988. Phonetic universals in consonant systems. In Larry M. Hyman and Charles N. Li, editors, *Language, Speech and Mind: studies in honour of Victoria A. Fromkin*. Routledge, London, pages 62–78.

Maddieson, Ian. 1978. Universals of tone. In Joseph H. Greenberg, editor, *Universals of Human Language (Phonology)*, volume 2. Stanford University Press, Stanford, Calif., pages 335–365.

Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge University Press, Cambridge, England.

Maddieson, Ian and Kristin Precoda. 1990. Updating upsid. *UCLA Working papers in Phonetics*, 74:104–111.

Ohala, John J. 1978. Production of tone. In Victoria A. Fromkin, editor, *Tone: a linguistic survey*. Academic Press, New York.

Pike, Kenneth. 1948. *Tone Languages*. The University of Michigan Press, Ann Arbor, US.

Redford, Melissa A., Chun Chi Chen, and Risto Miikkulainen. 2001. Constrained emergence of universals and variation in syllable systems. *Language and Speech*, 44:27–56.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. 1997a. The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25:255–286.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée, and Christian Abry.

1997b. Major trends in vowel system inventories. *Journal of Phonetics*, 25:233–253.

Stadler, Wolfram. 1988. *Multicriteria Optimization in Engineering and in the Sciences.* Plenum Press, New York.

Van Veldhuizen, David A. and Gary B. Lamont. 2000. Multiobjective evolutionary algorithms: Analyzing the state-of-the-art. *Evolutionary Computation*, 8(2):125–147. Special Issue: Multicriterion Optimization.

Vihman, Marilyn. 1977. *A Reference Manual and User's Guide for the Stanford Phonology Archive.* Stanford University. part I.

Wang, William S-Y. 1967. Phonological features of tone. *International Journal of American Linguistics*, 33:93–105.

Wang, William S-Y. 1968. The basis of speech. Project on linguistic analysis reports, University of California at Berkeley. Reprinted in The Learning of Language, ed. by C. E. Reed, 1971, and translated by Mieko Ogura, Tokyo: Kenkyusha, 1989.

Wang, William S-Y. 1976. Language change. *Annals of the New York Academy of Science*, 280:61–72.

Wang, William S-Y. 1991. Tone languages. In Kirsten Malmkjaer, editor, *The Linguistic Encyclopedia.* Routledge, London.