Evolutionary Discriminative Speaker Adaptation

Sid-Ahmed Selouani #1

# LARIHS Laboratory, Université de Moncton
218, Boul. J-D Gauthier, Shippagan (NB) E8S 1P6 Canada
1 selouani@umcs.ca

***Abstract*—This paper presents a new evolutionary-based approach that aims at investigating more solutions while simplifying the speaker adaptation process. In this approach, a single global transformation set of parameters is optimized by genetic algorithms using a discriminative objective function. The goal is to achieve accurate speaker adaptation whatever the amount of available adaptive data. Experiments using the ARPA-RM database demonstrate the effectiveness of the proposed method.**

I. Introduction

In linear transform speaker adaptation methods, a global transformation matrix is estimated in order to create a general model which better matches a particular target condition generated by a new speaker. To perform the adaptation on a small amount of data, a regression-tree-based classification is usually performed. The MLLR which is the most popular linear transform technique calculates a general regression transformation for each class, using data pooled within each class [1]. However, as mentioned in [2], transformation-based adaptation techniques suffer from two principal drawbacks. The first drawback is related to the fact that the type of the transformation function is fixed in advance to simplify the mathematical formalism. The second drawback lies in the bad asymptotic properties since these techniques may not achieve the level of accuracy obtained with speaker dependent systems even if the adaptation data quantity increases largely. The MAP-based techniques have better asymptotic properties but require more adaptation data compared to linear transform methods [2].

Over the last years, eigenvoice methods have become the backbone of most speaker adaptation methods. Eigenvoice modeling performs unsupervised and fast speaker adaptation through the use of eigen-decomposition, where the principal component analysis is used to project utterances of unknown speakers onto the orthonormal basis leading to speaker-dependent (SD) eigen coefficients.

Most conventional speaker adaptation approaches carry out an estimation of the linear transform parameters of mean and/or variance of speaker-independent (SI) HMMs. These parameters are used to perform the retraining by applying the maximum likelihood (ML) criterion to adjust the SI acoustic models so that they better fit the characteristics of a new speaker. Another recent and widely employed alternative approach consists of using discriminative linear transforms (DLT) to construct more accurate speaker adaptive speech recognition systems. Well-known discriminative criteria include maximum mutual information (MMI), minimum classification error (MCE), and minimum phone error (MPE) training. In [3], the MPE criterion is adopted for DLT estimation. Uebel and Woodland in [4] performed an interpolation of ML and MMI training criteria to estimate the DLT. In [5], Povey *et al.* studied the incorporation of the MAP algorithm into MMI and MPE for task and gender adaptations.

Many extensions have been proposed to improve the basic schemes of conventional and discriminative speaker adaptation which result in a wide range of hybrid approaches. In [6], Genetic Algorithms (GAs) have been used to enrich the set of SD systems generated by the eigen-decomposition. In a previous work, we have demonstrated the usefulness of GAs to optimize the MLLR based speaker adaptation [7]. In this paper, we extend our previous work by using a discriminative objective function instead of ML criterion to perform the speaker adaptation. Through the use of this evolutionary-based method, we expect to improve the accuracy of MLLR techniques. The rest of this paper is structured as follows. Section II gives background and technical details on the likelihood and discriminative based speaker adaptation methods. In Section III, we introduce and explain how to use GAs to optimize the discriminative speaker adaptation. Section IV presents and discusses the results obtained from comparing the proposed evolutionary-based system to baseline speaker adaptation systems. Finally, Section V concludes this work and gives some of its perspectives.

II. Discriminative Linear Transforms for Speaker Adaptation

A. Maximum Likelihood Based Adaptation

MLLR is a parameter transformation technique that has proven successful while using a small amount of adaptation data [1]. It computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. MLLR is a model adaptation technique that estimates a set of linear transformations for the mean (or variance) of Gaussian mixture HMM system. The effect of these transformations is to shift the component means in the initial system so that each state in the HMM is more likely to generate the adaptation data. The principle of mean transform in the MLLR scheme, assumes that Gaussian mean vectors are updated by linear transformation. Let μ_k be the baseline mean vector

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

and $\hat{\mu_k}$ the corresponding adapted mean vector for an HMM state k. The relation between these two vectors is given by: $\hat{\mu_k} = \mathbf{A}_k \xi_k$ where $\mathbf{A}_k$ is the $d \times (d+1)$ transformation matrix and $\xi_\mathbf{k} = [1, \mu_{k1}, \mu_{k2}, ..., \mu_{kd}]^t$ is the extended mean vector. It has been shown in [1] that maximizing the likelihood of an observation sequence o_t is equivalent to minimizing an auxiliary function Q given as follows:

$$Q = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(t)(o_t - \mathbf{A}_k \xi_k)^T C_k^{-1} (o_t - \mathbf{A}_k \xi_k), \quad (1)$$

where $\gamma_k(t)$ is the probability of being in the state k at time t, given the observation sequence o_t. C_k is the covariance matrix which is supposed to be diagonal. The general form for computing optimal elements of $\mathbf{A}_k$ is obtained by differentiating Q with respect to $\mathbf{A}_k$:

$$\sum_{t=1}^{T} \gamma_k(t) C_k^{-1} o_t \xi_k^t = \sum_{t=1}^{T} \gamma_k(t) C_k^{-1} A_k \xi_k \xi_k^t. \quad (2)$$

Depending on the amount of available adaptive data, a set of Gaussians, and more generally, a number of states will share a transform, and will be referred to as regression class r. Then, for a particular transform case $\mathbf{A}_k$, Gaussian components will be tied together according to a regression class tree and the general form of Equation 2 expands to:

$$\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{k_r}(t) C_{k_r}^{-1} o_t \xi_{k_r}^t = \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{k_r}(t) C_{k_r}^{-1} A_k \xi_{k_r} \xi_{k_r}^t. \quad (3)$$

In standard MLLR, the column by column estimation of $\mathbf{A}_k$ elements is given as follows:

$$a_i = G_i^{-1} z_i, \quad (4)$$

where z_i refers to the i^{th} column of the matrix which is produced by the left hand side of Equation 3, and where G_i is given by $\sum_{r=1}^{R} \hat{c}_{ii}^{(r)} \xi_{k_r} \xi_{k_r}^t$, where $c_{ii}^{(r)}$ is the i^{th} diagonal element of $\sum_{t=1}^{T} \gamma_{k_r}(t) C_{k_r}^{-1}$.

B. Discriminative Speaker Adaptation

Discriminative training algorithms, such as the MMI and MPE, have been successfully applied in large vocabulary speech recognition [8] and speaker adaptation tasks [3]. The main characteristic of these algorithms is that they consider not only the correct transcription of the training utterance, but also the competing hypotheses that are obtained by performing the recognition step.

In order to facilitate the inclusion of an evolutionary-based optimization, a baseline system is constructed by performing the MPE training to improve the SD acoustic models obtained by the MLLR. The SI model is first adjusted by MLLR using limited speaker-specific data. Then, the adapted SD model is updated by a MPE-based discriminative training. The numerator lattice is obtained through the alignment process on the transcriptions of the adaptation data. The denominator lattice is approximated with the N-best phone hypotheses after performing the recognition process on the adaptation data.

In the approach presented here, an MPE discriminative training is performed by using speaker-specific data. Many studies have demonstrated that MPE training outperforms MMI training [3]. Actually, MPE focuses on correctable errors in the training data rather than outliers which may reduce the effectiveness of MMI training. The MPE-based method consists of using a weak-sense auxiliary function in HMM to re-estimate the mean $\tilde{\mu}_{km}$ of mixture component m of state k of a new adapted model. This re-estimation is done as follows:

$$\tilde{\mu}_{km} = \frac{[\theta_{km}^{num}(O) - \theta_{km}^{den}(O)] + D_{km}\hat{\mu}_{km}}{[\gamma_{km}^{num} - \gamma_{km}^{den}] + D_{km}}, \quad (5)$$

where $\theta_{km}^{num}(O)$ and $\theta_{km}^{den}(O)$ are respectively the numerator and denominator sum of observation data weighted by the occupation probability for mixture m of state k; D_{km} is the Gaussian-specific smoothing constant; γ_{km}^{num} and γ_{km}^{den} are respectively the numerator occupation probabilities and the denominator occupation probabilities summed over time.

State-of-the-art techniques show that two different forms of discriminative speaker adaptation techniques (DSAT) are being used [9]. The first technique is based on ML speaker-specific transforms and its commonly used variant is the MLLR-based DSAT. In this approach, both ML-based and discriminative training are used. The MLLR-based adaptation is initially performed to produce a set of speaker-specific MLLR transforms. These transforms are then used to carry out the subsequent updates by using the MPE discriminative criterion. As stated by Raut *et al.* in [9] the use of ML-based speaker-specific transforms leads to more robustness to errors in the supervision hypothesis. The second approach is based on DLTs. In these DLT based methods, both of the transforms and the HMMs are estimated by using the MPE discriminative criterion. This yields a set of speaker-specific DLTs that are used for recognition. For the experiments presented in this paper, the MLLR-based DSAT is used.

III. Evolutionary Linear Transformation Paradigm

Genetic algorithms have been successfully integrated in the framework of speaker adaptation of acoustic models [7]. One of the approaches consists of using the genetic algorithm to enrich the set of speaker-dependent systems employed by the eigenvoices [6]. In this later work, the best results are obtained when the genetic algorithms are combined with the eigen decomposition. Since the eigen decomposition provides the weights of eigenvoices by using the EM algorithm, it can only find a local solution. In the GA-MPE-MLLR method presented here, the eigen-decomposition is avoided and the MPE criterion is used as an objective function. The MPE-based training has proven to be very effective in the generalization from training to test data, compared with the conventional maximum likelihood approach. The motivation for an evolutionary-based discriminative transform is based on the fact that DLTs were initially developed to correctly discriminate the recognition hypotheses for the best recognition results rather than just matching the model distributions.

In the GA-MPE-MLLR method, the mean transformation matrix (obtained by MLLR) provides the individuals involved in the evolutionary process. μ_k is the baseline mean vector and $\hat{\mu_k}$ is the adapted mean vector for an HMM state k. As seen above, the relationship between these two vectors is given by: $\hat{\mu_k} = \mathbf{A}_k \xi_k$. The $\mathbf{A}_k$ matrix will contain weighting factors that represent the individuals in an evolution process. These individuals evolve through many generations in a pool where genetic operators such as mutation and crossover are performed [10]. Some of these individuals are selected to reproduce according to their performance. The individuals' evaluation is performed through the use of the objective function. The evolution process is terminated when no improvement of objective function is observed. When the fittest individuals are obtained (the global optimized matrix $\mathbf{A_{gen}}$), they are used in the test phase to adapt data of new speakers. It is important to note that we do not need to determine the regression classes, since the optimization process is driven by a performance maximization whatever the amount of available adaptive data. The global GA-based adaptation process is illustrated by Figure 1. For any GA, a chromosome representation is needed to describe each individual in the population. The representation scheme determines how the problem is structured in the GA and also determines the genetic operators that are used. GA-MPE-MLLR involves genes that are represented by the components of $\mathbf{A_{gen}}$ matrix elements.

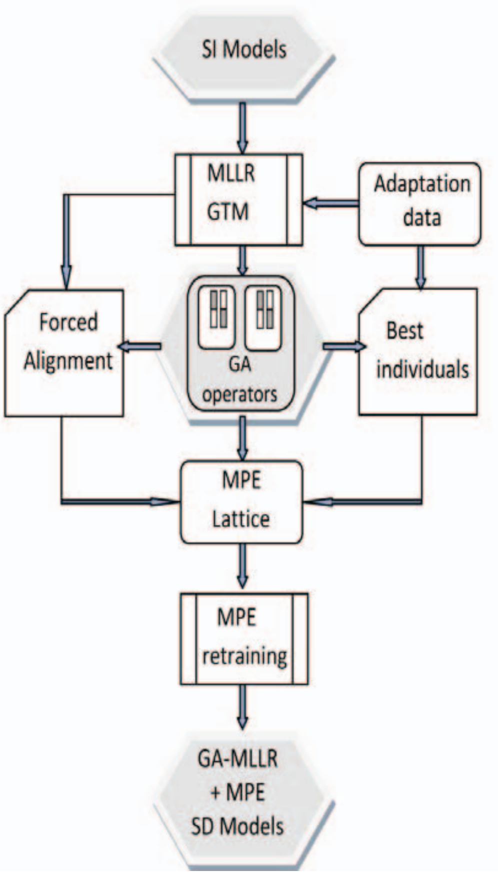

Fig. 1. Overview of the MPE-MLLR evolutionary-based speaker adaptation system.

A. Population Initialization

The first step to start the GA-MPE optimization is to define the initial population. This initial population is created by 'cloning' the elements of a global $\mathbf{A}$ matrix issued from a first and single MLLR pass. This procedure consists of duplicating the a_i elements of $\mathbf{A}$ to constitute the initial pool with a predetermined number of individuals. Hence, the pool will contain a_i^v individuals where v refers to an individual in the pool and it varies from 1 to $PopSize$ (population size). With this procedure, we expect to exploit the efficiency of GAs to explore the entire search space, and to avoid a local optimal solution. The useful representation of individuals involves genes or variables from an alphabet of floating point numbers with values varying within lower and upper bounds (b_1, b_2).

B. Objective Function

Formally, the optimization of the global transformation matrix requires finding the fittest individuals representing column vectors $a_i^v \in \mathcal{S}$, where $\mathcal{S}$ is the search space, so that a certain quality criterion is satisfied namely that objective function $\mathcal{F} : \mathcal{S} \to \mathcal{R}$ is maximized. $a_{i_{gen}}$ is the solution that satisfies:

$$a_{i_{gen}} \in \mathcal{S} \mid \mathcal{F}(a_{i_{gen}}) \geq \mathcal{F}(a_i^v) \qquad \forall a_i^v \in \mathcal{S}. \tag{6}$$

In the method we propose, the objective function (fitness) is defined in such a way that the newly genetically optimized parameters are guaranteed to increase the phone accuracy of adaptation data. For this purpose, we used the minimum phone error criterion utilizing phone lattices. The standard function reflecting the MPE criterion involves competing hypotheses represented as word lattices, in which phone boundaries are marked in each word to constrain the search during statistical estimation of an HMM model λ. For a specific model, this function is defined as:

$$F_{MPE}(\lambda) = \sum_{u=1}^{U} \sum_{s} P_\rho(s|O_u, \lambda) \sum_{q \in s} PhAcc(q), \tag{7}$$

where $P_l(s|O_u, \lambda)$ is the posterior probability of hypothesis s for utterance u given observation O_u, current model λ and acoustic scale ρ. $\sum_{q\in s} PhAcc(q)$, is the sum of phone accuracy measure of all phone hypotheses. The objective function used in the GA-MPE to evaluate a given individual a_i^v, considers the overall phone accuracy and then it is defined as:

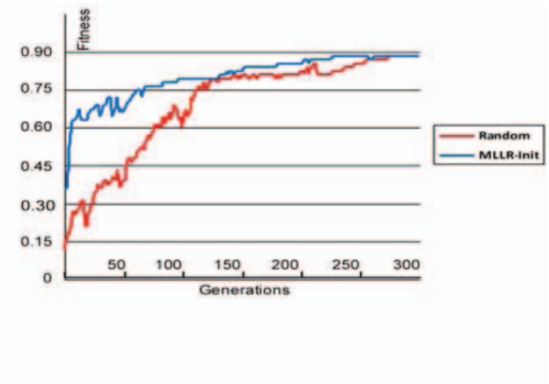

Fig. 2. Objective function variations of the ERS0 speaker, with random and basic MLLR initializations of population.

$$\mathcal{F}(a_i^v) = \sum_{\lambda} F_{MPE}(\lambda). \tag{8}$$

Objective function is normalized to unity. Figure 2 plots variations of the best individual $\mathcal{F}(a_{i_{gen}})$ with respect to the number of generations, in the case of totally random and first step MLLR initializations of population.

C. Selection Function

Since the offspring population is larger than the parent population, a mechanism has to be implemented in order to determine the individuals that will comply with the new parent population. The selection mechanism chooses the fittest individuals of the population and allows them to reproduce, while removing the remaining individuals. The selection of individuals to produce successive generations is based on the assignment of a probability of selection, P_v to each individual, v, according to its fitness value. In the 'roulette wheel' method [11], the probability P_v is calculated as follows:

$$P_v = \frac{\mathcal{F}(a_i^v)}{\sum_{k=1}^{PopSize} \mathcal{F}(a_i^k)} \tag{9}$$

where $\mathcal{F}(a_i^k)$ equals the value of objective function of individual k and $PopSize$ is the population size in a given generation. In the 'roulette wheel' variant implemented in GA-MPE, we introduced a dose of an *elitist* selection by incorporating in the new pool, the top two parents of previous populations to replace the two fitness-lowest offspring individuals [10].

D. Recombination

Recombination allows for the creation of new individuals (offspring) using individuals selected from the previous generation (parents). In the GA-MPE method, a combination of the conventional arithmetic crossover and guided crossover is used as a recombination operator. In the first step, this method selects thanks to the selection function, an individual as a first candidate ($cand_1$). A second candidate is then selected according to a quantity of what is called the mutual fitness $MF(X, cand_1)$ [12], where a choice for X as a second candidate is made if it maximizes the mutual fitness with the first candidate. The general computation of the mutual fitness is given by:

$$MF(A,B) = \frac{[\mathcal{F}(A) - \mathcal{F}(B)]^2}{Distance(A,B)^2} \tag{10}$$

The parents $cand_1$ and $cand_2$ are now selected and the convex combination can be applied according to the following equations :

$$\begin{cases} mix = (1 + 2 * \beta) * rand - \beta \\ x' = mix * cand_1 + (1 - mix) * cand_2 \\ y' = (1 - mix) * cand_1 + mix * cand_2, \end{cases} \tag{11}$$

where $rand$ is a Gaussian random value. If β is set to 0, the resulting crossover is a simple crossover. If β is set to a positive value this may increase the diversity of the individuals of the population and may allow children to explore the solution space beyond the domain investigated by their parents.

E. Mutation

Mutation operators tend to make small random changes on the individual components in order to increase the diversity of the population. Mutation consists of randomly selecting one gene x of an individual a_i^X and slightly perturbating it. In GA-MPE, the offspring mutant gene, $x^{''}$, is given by:

$$x'' = x + \mathcal{N}_k(0,1) \tag{12}$$

where $\mathcal{N}_k(0,1)$ denotes a random variable of normal distribution with zero mean and standard deviation 1 which is to be sampled for each component individually. The Gaussian-based alteration on the selected offspring individuals allows the extension of the search space and theoretically improves the ability to deal with new speaker related conditions.

F. Termination

The evolution process is terminated when a number of maximum generations is reached. No improvement of the objective function is observed beyond a certain number of generations. It is also important to note that as expected, the single class MLLR initialization yields a rapid fitness convergence, in contrast to the fully random initialization of the pool. When the fittest individual is obtained, it is used to produce a speaker-specific system from an (SI) HMM set.

IV. Experiments

A. Resources and Tools

The ARPA-RM database is used to evaluate the MPE-GA-MLLR technique. The (SI) subset of ARPA-RM is used for the training while a speaker dependent subset consisting of 47 sentences uttered by 6 speakers of ARPA-RM is used for the test [13]. The HTK toolkit implementing HMM-based speech recognition system is used throughout all experiments

TABLE I

COMPARISONS OF THE PERCENT WORD RECOGNITION ($\%C_{wrd}$) OF HMM-BASED CSR SYSTEMS FOR SELECTED DATA FROM THE ARPA-RM USED FOR TRAINING, ADAPTATION AND NEW SIX SPEAKERS FOR THE TEST.

Speaker	CMR0	DAS1	DMS0	DTB0	ERS0	JWSO
Unadapted	76.46	75.14	78.65	80.24	76.97	77.86
MLLR-ML	78.34	79.87	80.74	84.92	83.56	80.85
MLLR-DSAT	78.15	77.49	80.82	84.20	83.78	81.73
GA-MPE-MLLR	**79.71**	**81.30**	**82.14**	**86.26**	**84.75**	**83.92**

[14]. The adaptation was performed in unsupervised mode. The testing adaptation is performed with an enrollment set of 10 sentences. The acoustical analysis consists of 12 MFCCs which were calculated on a 30-msec Hamming window. The normalized log energy, the first and second derivatives are added to the 12 MFCCs to form a 39-dimensional vector. All tests are performed using 8-mixture Gaussian HMMs with triphone models.

B. Genetic Algorithm Parameters

To control the run behavior of a genetic algorithm, a number of parameter values must be defined. The initial population is composed of 200 individuals and is created by duplicating the elements of global transform matrix obtained after the first and single regression class MLLR. The genetic algorithm is halted after 350 generations. The percentage of crossover rate and mutation rate are fixed at 35% and 8% respectively. The number of total runs is fixed at 60. The GA-MPE-MLLR system uses a global transform where all mixture components are tied to a single regression class.

C. Result Discussion

Table I summarizes the word recognition rates obtained for the 6 speakers using four systems: the baseline HMMs-based system without any adaptation (unadapted) and using the ML criterion for recognition; the conventional MLLR using the ML criterion; the MLLR using a discriminative transformation (MLLR-DSAT) described in section II; and the system integrating the evolutionary subspace approach using the MPE criterion (GA-MPE-MLLR). The GA-MPE-MLLR leads to an improvement in the accuracy of word recognition rate reaching 8% compared to the baseline unadapted system and more than 3% compared to conventional MLLR. We have tested the fully random initialization of the population and the one using individuals cloned from MLLR global transformation matrix components. In both cases, the final performance is the same. However, the adaptation is reached rapidly (160 generations) with MLLR-based initialization.

V. CONCLUSION

The most popular approaches to speaker adaptation are based on linear transforms because they are considered more robust and use less adaptation data than the other approaches. This paper presented a framework demonstrating the suitability of genetic algorithms to improve unsupervised speaker adaptation using linear transforms. In fact, experiments show that the GA-MPE-MLLR approach outperforms the discriminative and conventional MLLR speaker adaptation technique. The main advantage of using the GA-based optimization is to avoid the regression class process usually done in conventional MLLR. Therefore the new speaker adaptation performance is not linked to the amount of available adaptive data. Many perspectives are open and may consist of fully automating the set up of genetic parameters. The ultimate objective is to give ASR systems auto-adaptation capabilities to face any acoustic environment change.

REFERENCES

[1] C. J. Legetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[2] C. Mokbel, “Online adaptation of hmms to real-life conditions: a unified framework,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 342–357, 2001.

[3] L. Wang and P. C. Woodland, “Mpe-based discriminative linear transforms for speaker adaptation,” *Computer Speech and Language*, vol. 22, no. 3, pp. 256–272, 2008.

[4] L. F. Uebel and P. C. Woodland, “Discriminative linear transforms for speaker adaptation,” in *Proceedings of ISCA ITRW Adaptation Methods for Automatic Speech Recognition*, Sophia-Antipolis, France, 2001, p. 6163.

[5] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland, “Mmi-map and mpe-map for acoustic model adaptation,” in *Proceedings of Eurospeech*, Geneva, Switzerland, Sept 2003, pp. 1891–1894.

[6] F. Lauri, I. Illina, D. Fohr, and F. Korkmazsky, “Using genetic algorithms for rapid speaker adaptation,” in *Proceedings of Eurospeech*, Geneva, Switzerland, Sept 2003, pp. 1497–1500.

[7] S. A. Selouani and D. O’Shaughnessy, “Speaker adaptation using evolutionary-based linear transform,” in *International Conference on Spoken Language Processing*, Pittsburgh, November 2006, pp. 1109–1112.

[8] D. Povey, “Discriminative training for large vocabulary speech recognition,” Ph.D. dissertation, Department of Engineering, Univ. of Cambridge, Cambridge, 2003. [Online]. Available: http://sites.google.com/site/dpovey/my-publications

[9] C. K. Raut, K. Yu, and M. J. F. Gales, “Adaptive training using discriminative mapping transforms,” in *Proceedings of Interspeech*, Brisbane, Australia, Sept 2008, pp. 1697–1700.

[10] Z. Michalewicz, *Genetic Algorithms + Data Structure = Evolution programs*, Springer-Verlag, Ed. New York: AI Series, 1996.

[11] D. E. Goldberg, *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley publishing, 1989.

[12] K. Rasheed and H. Hirsh, “Guided crossover: A new operator for genetic algorithm based optimization,” in *In Proceedings of the Congress on Evolutionary Computation*, Indianapolis, IN , USA, Apr 1997, pp. 1535–1541.

[13] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, “1000-word resource management data base for continuous speech recognition,” in *Proceedings of ICASSP*, New York, USA, Apr 1988, pp. 651–654.

[14] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, “Htk-hidden markov model toolkit,” 1995. [Online]. Available: http://htk.eng.cam.ac.uk/