SPEAKER ADAPTATION USING EVOLUTIONARY-BASED LINEAR TRANSFORM

Sid-Ahmed Selouani[1]*, and Douglas O'Shaughnessy*[2]

[1]Université de Moncton, Campus de Shippagan, E8S 1K9 NB, Canada
[2]INRS-EMT, Université du Québec, H5A 1C4, Montréal, Canada
selouani@umcs.ca, dougo@inrs-emt.uquebec.ca

Abstract

This paper presents a technique to adapt HMMs to new speakers by using Genetic Algorithms (GAs) in unsupervised mode. The implementation requirements of GAs, such as genetic operators and objective function, have been chosen in order to give more reliability to a global linear transformation matrix. By implementing a 'survival of the fittest' strategy, the proposed GA-MLLR approach allows to maintain and manipulate a population of a wide range of solutions.Experiments have been performed on ARPA-RM and TIMIT databases using a triphones HMM-based system. Results show that from a new speaker, significant decrease of word error rate which can reach 6% for a particular speaker, has been achieved by the evolutionary approach, compared to the conventional MLLR-based adaptation method.

Index Terms: speaker adaptation, genetic algorithms, linear transform, robustness.

1. Introduction

Speaker adaptation remains one of the fundamental challenges facing the development of speech recognition technology. Even if Continuous Speech Recognition (CSR) systems have achieved increasingly high recognition accuracy in a speaker-dependant context (SD), their performance often degrades when there are mismatches between training and testing conditions, particularly those introduced by new speakers. If a large amount of test data is available, a simple re-training technique can be used. However, in real-world applications it is hard to acquire a large amount of training data from a new test speaker. Reliable and efficient systems seek to achieve both fast and unsupervised adaptation by using a small amount of data. Among the wide variety of speaker adaptation techniques, Maximum Likelihood Linear Regression (MLLR) remains one of the most popular one [4]. In conventional MLLR, a global transformation matrix is estimated in order to make a general model better match a particular target condition. To permit adaptation on a small amount of data a regression-tree-based classification is performed. The MLLR framework calculates a general regression transformation for each class, using data pooled within each class. However, as mentioned in [7], transformation-based adaptation techniques such as MLLR suffer from two principal drawbacks. First, the type of transformation function used to reach the targeted model, is fixed in advance for only mathematical convenience. Second, these techniques do not have good asymptotic properties in the sense that the speaker adaptive technique may not achieve the level of accuracy obtained with the speaker dependent system even if the quantity of target data increases largely. In recent years, the conventional MLLR has been extended to perform unsupervised and fast speaker adaptation through the use of the eigen-based MLLR, where the principal component analysis (PCA) is used to project utterances of unknown speakers onto the orthonormal basis leading to SD eigen coefficients [1]. In [3], Genetic Algorithms (GAs) have been used to enrich the set of SD systems generated by the eigen-decomposition.

In the present work we propose an approach that aims to investigate more solutions while simplifying the adaptation process through the determination of a single global transformation set of genetically optimized parameters. The goal is to achieve adaptation, whatever the amount of available adaptive data, and to overcome the problem of huge memory and time consuming requirements of the eigen-MLLR techniques. The rest of this paper is organized as follows. In section 2 we proceed with an overview of the proposed GA-MLLR method. Section 3 describes the evolutionary-based paradigm that we introduce to perform speaker adaptation. Section 4 is devoted to the experimental part by validating the proposed algorithm for an unsupervised adaptation. Finally, in section 5, we conclude and discuss possible perspectives of this work.

2. Overview of the method

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 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 [4] that maximizing the likelihood of an observation sequence o_t is equivalent to minimizing an auxiliary function Q given as follows:

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

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 $\mathbf{A}_k$:

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

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 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:

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

where z_i refers to the i^{th} column of the matrix which is produced by the left hand side of 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 $\sum_{t=1}^{T}\gamma_{k_r}(t)C_{k_r}^{-1}$.

In the GA-MLLR method we propose, the $\mathbf{A}_k$ matrix will contain weighting factors that represent the individuals in the evolution process. These individuals evolve through many generations in a pool where genetic operators such as mutation and crossover are applied [2]. Some of these individuals are selected to reproduce according to their performance. The individuals' evaluation is performed through the use of an objective function (*fitness*). The evolution process is terminated when no improvement of objective function is observed. When the fittest individual is obtained, that is the global optimized matrix $\mathbf{A_{gen}}$; it is used in the test phase to adapt data of new speaker. Note that the problem of determining regression classes is not needed, since the optimization process is driven by a performance maximization whatever the amount of available adaptive data. The GA-based adaptation process is illustrated by Figure 1.

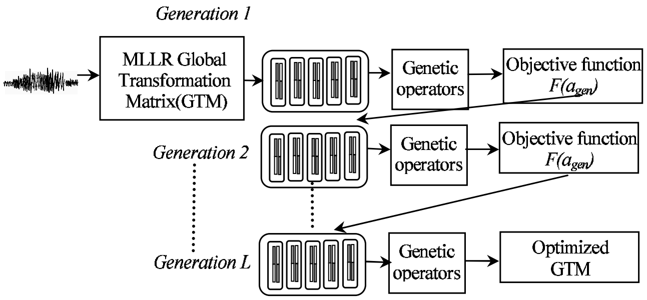

Figure 1: Overview of evolutionary-based linear transform.

3. Evolutionary Linear Transformation Paradigm

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 [2]. GA-MLLR involves genes that are represented by the components of $\mathbf{A_{gen}}$ matrix elements.

3.1. Population initialization

The first step to start the GA-MLLR optimization is to define the initial population pool. Contrary to common methods that randomly generate solutions for the entire population, we suggest the creation of an initial population by 'cloning' the elements of a global $\mathbf{A}$ matrix issued from a first and single MLLR pass. This procedure consists of duplicating the $a_i = G_i^{-1}z_i$ 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, in contrast to the random initialization, we expect to exploit the efficiency of GAs to explore the entire search space, and to avoid a local optimal solution.

3.2. Objective function

Formally, the optimization of the global transformation matrix requires finding 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 $\mathcal{F} : \mathcal{S} \rightarrow \mathcal{R}$ is maximized. $a_{i_{gen}}$ is the solution that satisfies:

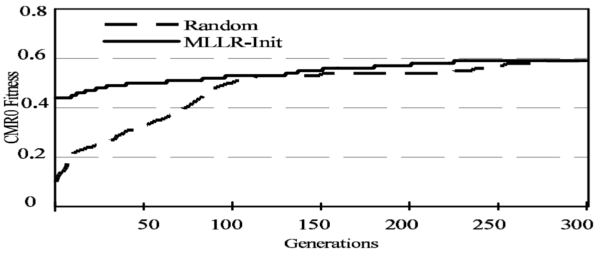

Figure 2: Objective function variations with random and basic MLLR initializations of population.

$$a_{i_{gen}} \in \mathcal{S} \mid \mathcal{F}(a_{i_{gen}}) \geq \mathcal{F}(a_i^v) \qquad \forall a_i^v \in \mathcal{S}. \tag{5}$$

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 a variant of the minimum phone error criterion (MPE) known as a very efficient discriminative training procedure, utilizing phone lattices [5]. 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:

$$F_{FPE}(\lambda) = \sum_{u=1}^{U} \sum_{s} P_l(s|O_u, \lambda) \sum_{q \in s} PhAcc(q), \tag{6}$$

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 l. $\sum_{q \in s} PhAcc(q)$, is a the sum of phone accuracy measure of all phone hypotheses. The objective function used in the GA-MLLR to evaluate a given individual a_i^v, considers the overall phone accuracy and then it is defined as:

$$\mathcal{F}(a_i^v) = \sum_{\lambda} F_{FPE}(\lambda). \tag{7}$$

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.

3.3. Selection function

Since the offspring population is larger than the parent population, a mechanism has to be implemented that allows to determine which individuals will conform to the new parent population. The selection mechanism chooses the fittest individuals of the population and allows them to reproduce, while killing off the other 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 [2], the probability P_v is calculated as follows:

$$P_v = \frac{\mathcal{F}(a_i^v)}{\sum_{k=1}^{PopSize} \mathcal{F}(a_i^k)} \tag{8}$$

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-MLLR, we introduced a dose of an *elitist* selection by incorporating in the new pool, the top two parents of previous population to replace the two fitness-lowest offspring individuals.

3.4. Recombination

Recombination allows for the creation of new individuals based on previous generation. Many schemes of recombination exist and are being used in GAs [6]. In GA-MLLR method, heuristic crossover is chosen to be used as a recombination operator. It generates a random number from a uniform distribution and does an exchange of the parents' genes x and y belonging to a_i^X and a_i^Y individuals respectively, on the offspring genes ($x^{'}$ and $y^{'}$). The choice of this type of crossover is justified by the fact that it is the only operator that utilizes fitness information. The offspring is created using the following equation:

$$\begin{aligned} x' &= x + U(0,1)(x - y) \\ y' &= x, \end{aligned} \tag{9}$$

where a_i^X is assumed to have better fitness than a_i^Y. $U(0,1)$ is a random variable of uniform distribution on the interval (0,1).

3.5. Mutation

Mutation operators tend to make small random changes in an attempt to explore all regions of the solution space. Mutation consists of randomly selecting one gene x of an individual a_i^X and slightly perturbating it. In GA-MLLR, the offspring mutant gene, $x^{''}$, is given by:

$$x'' = x + \mathcal{N}_k(0,1) \tag{10}$$

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 unseen speaker related conditions.

3.6. Termination

The evolution process is terminated when a number of maximum generations is reached. This maximum number of generations is obtained in such a way that it corresponds to the stabilization of objective function. As shown in Figure 2, 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.

4. Experiments

In the following experiments the TIMIT and ARPA-RM databases were used to evaluate the GA-MLLR technique. HTK toolkit, the HMM-based speech recognition system, has been used throughout all experiments. The *dr1* subset from the TIMIT database was chosen for the training while a speaker dependant subset of ARPA-RM consisting of 47 sentences of ARPA-RM uttered by 6 speakers is used for the test. The adaptation was performed in unsupervised mode. The testing adaptation is performed with an enrollment set of 10 sentences. All speech was coded into frames consisting 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 were performed using 8-mixture Gaussian HMMs with tri-phone models.

To control the run behaviour of a genetic algorithm, a number of parameter values must be defined. The initial population is composed of 150 individuals and was created by duplicating the elements of global transform matrix obtained after the first and single regression class MLLR. The genetic algorithm was halted after 300 generations. The percentage of crossover rate and mutation rate are fixed respectively at 35% and 8%. The number of total runs was fixed at 50. GA-MLLR uses a global transform where all mixture components were tied to a single regression class. Table 1 summarizes the word recognition rates obtained for the 6 speakers using 3 systems: the baseline HMMs-based CSR system without any adaptation (unadapted); the conventional MLLR and a GA-MLLR. GA-MLLR provides an improvement in the accuracy of word recognition rate varying from 2% to 6% compared to conventional MLLR. We have tested the fully random initialization of population and the one using individuals cloned from MLLR global transformation matrix components. For both cases, the final performance is the same. However, the adaptation is reached rapidly (180 generations) with MLLR-based initialization.

Table 1: Comparisons of the percent word recognition ($\%C_{wrd}$) of HMM-based CSR systems for selected data from the ARPA-RM used for adaptation and test, while the TIMIT $dr1$ subset was used for training.

Speaker	CMR0	DAS1	DMS0	DTB0	ERS0	JWSO
unadapted	76.12	74.23	78.38	79.21	77.46	77.82
MLLR	79.88	79.33	81.08	85.33	82.78	81.15
GA-MLLR	83.90	82.75	86.45	89.11	84.48	87.89

5. Conclusion

This work has illustrated the suitability of an evolutionary-based technique to adapt speech uttered by new speakers. Experiments show the effectiveness of GA-MLLR compared to conventional MLLR in rapid and unsupervised speaker adaptation. Using this approach avoids the regression class process and hence the performance is not linked to the amount of available adaptive data. Investigating more solutions while simplifying the adaptation process through the use of only one global transformation set of genetically optimized parameters, permits us to overcome the problem of huge memory and time consuming requirements such as in the eigen-MLLR-based techniques. Present goals of our work consist of fully automating the set up of genetic parameters. The ultimate objective is to give CSR systems auto-adaptation capabilities to face any acoustic environment change.

6. References

[1] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, “Fast speaker adaptation using eigenspace-based maximum likelihood linear regression”, Proc. ICSLP, Vol. 3, pp. 742–745, 2000.

[2] L. Davis, *The genetic algorithm handbook*, Ed. New York: Van Nostrand Reinhold, ch.17, 1991.

[3] F. Lauri, I. Illina, D. Fohr, F. Korkmazsky “Using genetic algorithms for rapid speaker adaptation”, Proc. Eurospeech, pp. 1497-1500, 2003.

[4] 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.

[5] L. Wang, and P.C Woodland, “MPE-based discriminative linear transform for speaker adaptation”, Proc. IEEE-ICASSP, Vol. I, pp. 321–324, 2004.

[6] Z. Michalewicz, *Genetic Algorithms + Data Structure = Evolution programs,* AI series. Springer-Verlag, New York, 1996.

[7] C. Mokbel, “Online adaptation of HMMs to real-life conditions: a unified framework”, IEEE-trans. on SAP, Vol.9, No 4, pp. 342-357, May 2001.