Dynamic Estimation of Phoneme Confusion Patterns with a Genetic Algorithm to Improve the Performance of Metamodels for Recognition of Disordered Speech

Santiago Omar Caballero Morales and Felipe Trujillo Romero

Technological University of the Mixteca, UTM, Highway to Acatlima, Km. 2.5, Huajuapan de Leon, Oaxaca, 69000
{scaballero,ftrujillo}@mixteco.utm.mx

Abstract. A field of research in Automatic Speech Recognition (ASR) is the development of assistive technology, particularly for people with speech disabilities. Diverse techniques have been proposed to accomplish accurately this task, among them the use of Metamodels. In this paper we present an approach to improve the performance of Metamodels which consists in using a speaker's phoneme confusion matrix to model the pronunciation patterns of this speaker. In contrast with previous confusion-matrix approaches, where the confusion-matrix is only estimated with fixed settings for language model, here we explore on the response of the ASR for different language model restrictions. A Genetic Algorithm (GA) was applied to further balance the contribution of each confusion-matrix estimation, and thus, to provide more reliable patterns. When incorporating these estimates into the ASR process with the Metamodels, consistent improvement in accuracy was accomplished when tested with speakers of mild to severe dysarthria which is a common speech disorder.

Keywords: Genetic Algorithms, Disordered Speech Recognition, Metamodels.

1 Introduction

Dysarthric speech is different from normal speech as it is affected by breathing and articulation abnormalities which cause performance in Automatic Speech Recognition (ASR) to decrease considerably [9,10,13,15,18]. These abnormalities decrease the speaker's intelligibility and restrict the speaker's phonemic repertoire, thus some sounds or phonemes cannot be uttered or articulated correctly. In ASR this leads to an increase of deletion, insertion, and substitution of phonemes [8,14,15,16].

Most speaker adaptation algorithms are based on the principle that it is possible to apply a set of transformations to the parameters of the acoustic models of an ASR system to move them closer to the voice of an individual [20]. Whilst this

I. Batyrshin and M. González Mendoza (Eds.): MICAI 2012, Part II, LNAI 7630, pp. 175–187, 2013.

has been shown to be successful for normal speakers, it is less successful in cases where phonatory dysfunction is present (and the phoneme uttered is not the one that was intended but is substituted by a different phoneme or phonemes) as often happens in dysarthric speech [7].

In [3] it was proposed that, instead of adapting the acoustic models of the ASR system, the errors made by the speaker at the phonetic level could be modelled to attempt to correct them. This was a concept that was previously proposed in [11] to correct the phoneme output of an ASR system, and in [3] this was further extended to accomplish improvement at the word recognition level. This approach, which made use of a phoneme-confusion matrix to get the estimates of the speaker's pattern of phonetic errors, was also explored in [12], [19], and [17].

In the field of artificial intelligence, a confusion-matrix is a visualization tool typically used in supervised learning. Each column of the matrix represents the instances in a recognised class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (e.g., commonly mislabelling or classifying one as another).

As implementation techniques to incorporate the phoneme confusion-matrix estimates into the ASR process are Weighted Finite-State Transducers (WFSTs) [3,11,17], and Hidden Markov Models (HMMs, Metamodels) [3,12]. However an important issue has remained within these studies, which is the ASR's phoneme output used to estimate the phoneme confusion-matrix.

From the mathematical model of the ASR process:

$$\hat{W} = \max_{W \in L} P(O|W)P(W) \tag{1}$$

the most likely sequence of words $\hat{W}$ given some acoustic observation O can be estimated as the product of two probabilities: $P(W)$, the *prior probability*, which is obtained from the Language Model (L); and $P(O|W)$, the *observation likelihood*, which is obtained from the acoustic model.

$P(W)$ is usually estimated by using N-gram grammars, and $P(O|W)$ is usually modelled by using Hidden Markov Models (HMMs), or Artificial Neural Networks (ANN). For word recognition tasks, the ASR system estimates the sequence of the most likely phoneme models (HMMs) that represent the speech O. Then, this sequence is restricted to form words by incorporating a lexicon. Finally, these words are restricted by the information of the language model $P(W)$ to form valid sequences of words.

The influence of $P(W)$ is very important for the final result, and hence, of the output generated for confusion-matrix estimation. The performance of the implementation techniques (HMMs, WFSTs) [3,12,17] rely critically on the accuracy of the phoneme output sequences used for confusion-matrix estimation [4]. In the studies using the confusion-matrix modelling approach, usually a single confusion-matrix is used for training purposes. In [11] the phoneme sequences were obtained with unrestricted phoneme language. In [4] the training phoneme sequences were obtained from the phonetic transcriptions of the word output of

the ASR system. It was found that these sequences were more accurate than those obtained directly at the phonetic level (with a phoneme language model).

However, with this approach, high accuracy in phoneme recognition does not always correlate with high accuracy in word recognition. This is because of the different language models restrictions used for confusion-matrix estimation and word recognition evaluation. Another issue is related to the problem observed when the data available for confusion-matrix estimation is small, which leads to poor estimates, and in practice, this is the normal situation.

In this paper, we propose an extended approach to obtain better estimates of a phoneme confusion-matrix to increase ASR performance for dysarthric speech. Instead of just considering a confusion-matrix estimated with a single language model (either phoneme-based or word-based), we consider a dynamic estimation of diverse confusion-matrices with different language model restrictions. These matrices then are weighted and integrated into a single confusion-matrix, which would have more information about the behaviour of the speaker's confusion patterns across different language model scenarios. The implementation technique for the incorporation of the confusion-matrix into the ASR process is the extended version of the Metamodels (discrete HMMs) [3,12] which was presented in [2]. The weights to balance the contribution of each confusion-matrix are estimated by means of a Genetic Algorithm (GA), which also performs optimization on the structure of the extended Metamodel. When evaluating this approach with a well known database of dysarthric speech, performance was significantly higher when compared with the single confusion-matrix estimation approach, and also, when compared with a Speaker Adaptive (SA) ASR system.

Hence, this paper is structured as follows: in Section 2 information about the integration of the confusion-matrix estimates into the ASR process is reviewed. Then, we present the details of the dynamic estimation of the confusion-matrices. In Section 3 the extended Metamodel for implementation is reviewed, and then in Section 4 the Genetic Algorithm for optimization of the contributions of the dynamic confusion-matrices and the Metamodel is presented. In Section 5 the details of the experiments are presented, which involves the training and evaluation procedures. Finally, in Section 6, the results and future work are discussed.

2 Phoneme Confusion-Matrix as Resource for Error Correction

In Figure 1 an example of a phoneme confusion-matrix is shown, where rows represent the phonemes intended or uttered by the speaker (Stimulus), and the columns represent the decoded phonemes given by the ASR system (Response). The classification of phonemes to estimate a phoneme confusion - matrix is performed by the alignment of two phoneme strings (or sequences):

- P, the reference (correct) phoneme transcription of the sequence of words W uttered by a speaker.
- $\tilde{P}^*$, the sequence of phonemes decoded by the ASR system.

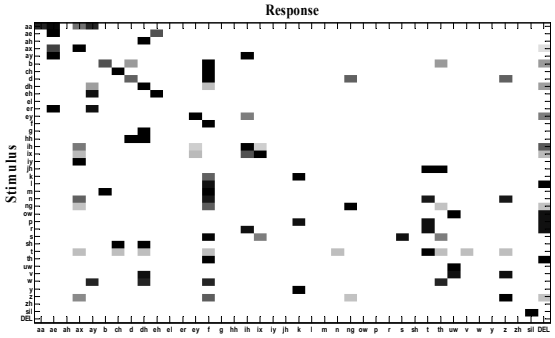

Fig. 1. Example of a phoneme confusion-matrix

As $\tilde{P}^*$ is the system's output, it might contain several errors. Based on the classification performed by the aligner, these are identified as substitution (S), insertion (I), and deletion (D) errors. Thus, the performance of ASR systems is measured based on these errors, and two metrics are widely used for phoneme and word ASR performance: Word Accuracy (WAcc) $= \frac{N-D-S-I}{N}$, and the Word Error Rate(WER) $= 1 - WAcc$. Where N is the number of elements (words or phonemes) in the reference string (P). Thus, the objective of the statistical modelling of the phoneme confusion-matrix is to estimate W from $\tilde{P}^*$. This can be accomplished by the following expression [3]:

$$W^* = \max_{P} \prod_{j}^{M} Pr(p_j) Pr(\tilde{p}_j^* | p_j) \tag{2}$$

where p_j is the j'th phoneme in the postulated phoneme sequence P, and $\tilde{p}_j^*$ the j'th phoneme in the decoded sequence $\tilde{P}^*$ (of length M). Eq. 2 indicates that the most likely word sequence is the sequence that is most likely given the observed phoneme sequence from a speaker. The term $Pr(\tilde{p}_j^* | p_j)$ represents the probability that the phoneme $\tilde{p}_j^*$ is recognized when p_j is uttered, and is obtained from a speaker's confusion-matrix. This element is integrated into the recognition process as presented in Figure 2.

This information then can be modelled by techniques to improve the baseline ASR's output. Evaluation is performed when $\tilde{P}^*$ (which now is obtained from test speech) is decoded by using the "trained" techniques into sequences of words W^*. The correction process is done at the phonetic level, and by incorporating a word-language model a more accurate estimate of W is obtained. In this work, the extended Metamodels are the technique used for the modelling of the confusion-matrix and implementation fo the ASR process. This is presented in Section 3.

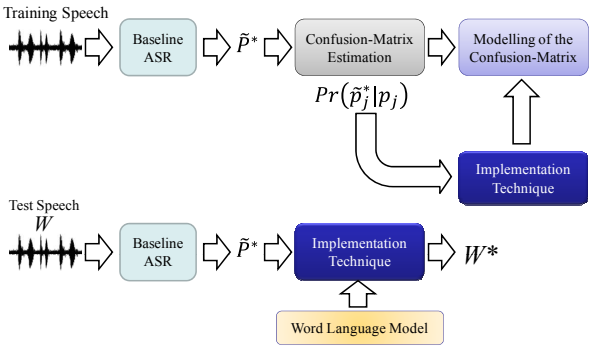

Fig. 2. Training and testing process of the confusion-matrix approach

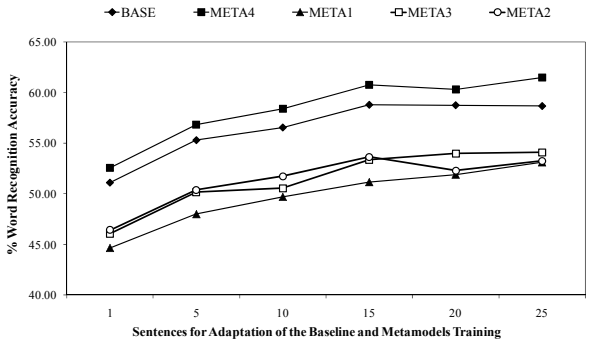

Fig. 3. Comparison of performance of Metamodels trained with different phoneme sequences

2.1 Dynamic Estimation of the Phoneme Confusion-Matrix

As presented in Section 1, the studies that have used the phoneme confusion-matrix as resource for error modelling have only used a single baseline condition to estimate the confusion patterns. This is, a single language model restriction (either phoneme-based or word-based). We consider necessary to study the behaviour of the ASR for different language model restrictions as this affects the phoneme/word output, which is the base for confusion-matrix estimation.

For example, consider Figure 3, which shows the performance of the single phoneme confusion-matrix approach when trained and tested with different number of phoneme sequences generated with different language model restrictions. These results were obtained during a previous exploratory study with the Nemours database of dysarthric speech [1], which is presented in Section 5. The baseline (BASE) represents the adapted word output from a standard ASR system (with continuous HMMs for acoustic modelling). META1 represents the performance of the implementation technique of the Metamodels (see Section

3) trained with the confusion-matrix estimated from phoneme sequences generated with a phoneme-based (PB) language model with no probability restrictions [1](s=0). META2 and META3 are the performance of the same Metamodels with different levels of restrictions for the PB language model (s=10 and s=25). META4 is the performance when the confusion-matrix is estimated from the phoneme transcriptions of the word output generated with a highly restricted word-based (WB) language model (i.e., s >30). As it is presented, there are significant differences in performance.

Also it was observed that there was variability in the patterns of phoneme confusions observed with different levels of language model restrictions. In Figure 4 is shown a sequence of confusion-matrices estimated from phoneme sequence obtained with different grammar scale factors for the language model. In the matrix estimated from phoneme sequences obtained from the transcriptions of words obtained with a WB language model, there are missing confusion patterns that are present in the matrices obtained with a PB language model with different restrictions (s=1,5). These estimates are considered to be important, especially when training data is sparse and estimation of unseen data is required. These confusion patterns potentially can be present in the unseen test set.

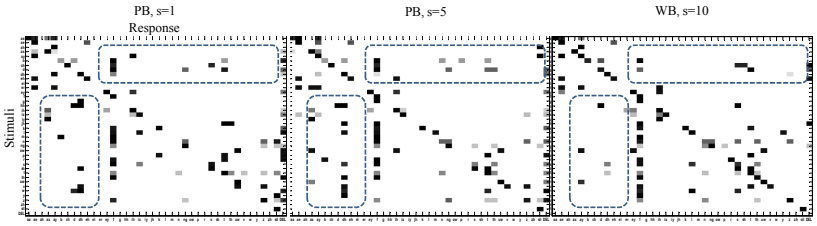

Fig. 4. Comparison of phoneme confusion-matrices obtained with different language model conditions

By considering this information, a more complete confusion-matrix can be estimated. Note however that, incorporating all this information in a single confusion-matrix is not an easy task. This is because some patterns do not contribute in the same way as others. For example, if just a number of matrices are estimated and averaged to get a single matrix, the most dominant patterns across all the matrices will be reinforced in the single matrix. On the other hand, the less dominant, but still important, will be reduced even further. This was considered to be a problem similar to the mixture of gaussian distributions where, in this case, would be a mixture of phoneme confusion-matrices.

We start by taking as reference the definition of the *emission probabilities* of the HMMs, $B = \{b_i(\mathbf{o}_t)\}$, where each term represents the probability of an observation vector $\mathbf{o}_t$ being generated from a state j:

[1] The grammar scale factor is a variable used to set the influence (or restriction) of the language model on the ASR process[20] and is identified as $s \in (0, 30)$.

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} C_{jk} N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{3}$$

In Eq. 3, K denotes the number of mixture-components, C_{jk} is the weight for the k-th mixture component satisfying $\sum_{k=1}^{K} C_{jk} = 1$, and $N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ denotes a single Gaussian density function with mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix $\boldsymbol{\Sigma}_{jk}$ for state j.

For this work, K would be the number of confusion-matrices to consider, each one obtained from a different language model condition. C_k the weight which would measure the contribution, or importance, of each k-th confusion-matrix. And $Pr(\tilde{p}_j^*|p_j)^k$ the discrete distribution probabilities associated to the k-th phoneme confusion-matrix (see Eq. 2). Hence, the estimated dynamic phoneme confusion-matrix can be expressed as:

$$Pr(\tilde{p}_j^*|p_j)_{dyn} = \sum_{k=1}^{K} C_k Pr(\tilde{p}_j^*|p_j)^k \tag{4}$$

In Section 4 the algorithm designed to estimate the weights C_k is presented.

3 Extended Metamodels

The Metamodels are a technique proposed to model a speaker's confusion-matrix $Pr(\tilde{p}_j^*|p_j)$ and to incorporate this information into the ASR process [3,12]. These consist of discrete HMMs with the structures shown in Figure 5, which allow the modelling of insertion and deletion patterns. Each state of a metamodel has a discrete probability distribution over the symbols for the set of phonemes (i.e., $Pr(\tilde{p}_j^*|p_j)$), plus an additional symbol labelled DEL (deletion). The central state (C) of a metamodel for a certain phoneme models correct decodings, substitutions and deletions of this phoneme made by the ASR system. States BF and AT model (possibly multiple) insertions before and after the phoneme.

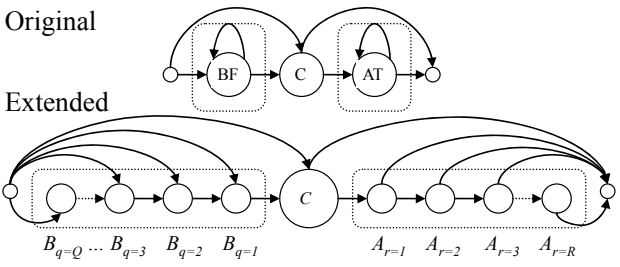

Fig. 5. Original and Extended Metamodels

The extended version of the Metamodels was taken for this work as it was reported to perform better than the original version [2] for the modelling of insertion patterns. Hence, the difference to the original Metamodels is the extension of the insertion states to model the insertion-context associated to each phoneme: the states B_q model the q-th *insertion-before* a phoneme, and A_r the r-th *insertion-after* a phoneme. These $q = 1 \dots Q$ and $r = 1 \dots R$ indexes identify the contexts of such insertions, where Q and R represent the length of the contexts. More details about this technique can be found in [2].

4 Optimization Method: Genetic Algorithm

A Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution and generates useful solutions to optimization problems [6]. In a GA, the solutions for an optimization problem receive a penalization score based on their quality or "fitness", which determines their opportunities for reproduction. It is expected that parent solutions of very good quality will produce offsprings (by means of reproduction operators such as crossover or mutation) with similar or better characteristics, improving their fitness after some generations. For our problem, we need to find the following:

- a vector of values C_k for the weights of the K phoneme confusion-matrices used to dynamically obtain a representative matrix for error modelling;
- the grammar scale factor s more suitable to obtain representative phoneme confusion-matrices;
- the length of the insertion-contexts Q and R for the extended Metamodels.

The coding scheme and GA implementation are discussed in the following sections.

4.1 Chromosome Representation and Fitness Evaluation

The chromosome of the GA is presented in Figure 6. The first 10 elements (genes) consist of the weights of C_k, where each one is estimated initially as $C_k = rand(1, 10)/10$ (hence, K=10 phoneme confusion-matrices are considered). Note that these values are not integers, and that the initial set of values for C_k may not sum to 1.0 because each one is estimated individually. Thus, these values must be normalized after the last value is estimated.

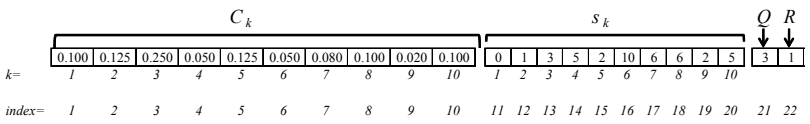

Fig. 6. Chromosome representation of the parameters for the Metamodels and the estimation of the dynamic confusion-matrix

From the 10 phoneme confusion-matrices to be used, the first 9 are estimated from the ASR system's phoneme sequences obtained with a phoneme-based (PB) language model. The last one is estimated from the transcription of the word output of the same ASR system, which is obtained with a word-based (WB) language model. Thus, if the following 10 genes correspond to the grammar scale factors $s_k \in(0,20)$ (which have integer values) for the estimation of each k-th confusion matrix, the first 9 are for the PB language model, and the last one for the WB language model. Finally, in the last two genes (21, 22) are stored the lengths of the insertion-contexts Q and R ($\in$ (0,10)).

The initial population for the GA consists of 10 individuals, where the first element is the extended metamodel built with the single confusion-matrix approach (hence, genes 1 - 10, 11-20, and 21-22 remain constant with their original values), and the remaining individuals are randomly generated within the range of values specified above.

The fitness value was measured as the Word Recognition Accuracy (%WAcc) obtained by the built Metamodel with the single or dynamic confusion-matrix approach on a "control" speech set (see Section 5).

4.2 Operators: Selection and Reproduction

- **Selection**: How to choose the eligible parents for reproduction was based on the Roulette Wheel and was implemented as follows:
  1. For each of the 10 best individuals in the population, compute its fitness value.
  2. Compute the selection probability for each x_i individual as: $p_i = \frac{f_i}{\sum_{k=1}^{H} f_k}$, where H is the size of the population (sub-set of 10 individuals), and f_i the fitness value of the individual x_i.
  3. Compute the accumulated probability q_i for each individual as: $q_i = \sum_{j=1}^{i} p_j$.
  4. Generate a uniform random number $r \in \{0, 1\}$.
  5. If $r < q_i$, then select the first individual (x_1), otherwise, select x_i such that $q_{i-1} < r \leq q_i$.
  6. Repeat Steps 4 and 5 H times (until all H individuals are selected).
- **Crossover**: Uniform crossover was used for reproduction of parents chosen by the Roulette Wheel method. A template vector of dimension 1×22 was used for this, where each of its elements received a random binary value (0, 1). Offspring 1 is produced by copying the corresponding genes from Parent 1 where the template vector has a value of 0, and copying the genes from Parent 2 where the template vector has a value of 1. Offspring 2 is obtained by doing the inverse procedure. 10 offsprings are obtained by crossover from the 10 individuals in the initial population. This increased the size of the population to 20.
- **Mutation**: The mutation scheme consisted in randomly changing all chromosome values in the best 5 individuals. In this way, a population of 25 individuals is obtained for the GA.
- **Stop Condition**: The algorithm was performed for 20 iterations as it was observed that convergence was stable by that point.

5 Experiments on Dysarthric Speech

5.1 Speech Data and Baseline Recogniser

For the experiments with dysarthric speech, the Nemours database [1] was used. This database consists of a collection of 814 short sentences spoken by 11 american-english speakers (74 sentences per speaker) with varying degrees of dysarthria resulting from either Cerebral Palsy or head trauma (data from only 10 speakers was used as some data is missing for one speaker). The sentences are nonsense phrases that have a simple syntax of the form "the X is Y the Z", where X and Z are monosyllabic nouns (74 in total) and Y is a bisyllabic verb (37 in total) in present participle form (for instance, the phrases "The shin is going the who", "The inn is heaping the shin", etc.).

In addition, data from a reference speaker with normal speech is included. This speaker utters each one of the 740 sentences spoken by the dysarthric speakers. Hence, a baseline speech recogniser was built with this speaker's data. This task was accomplished with the HTK Toolkit [20]. In general, 39 monophone acoustic HMMs were constructed with a standard three state left-right topology with eight mixture components per state. The front-end used 12 MFCCs plus energy, delta and acceleration coefficients. A frame period of 10 msec with a Hamming window of 25 msec and 26 filter-bank channels were used.

To adapt the baseline system to each dysarthric speaker, a common practice when using commercial ASR systems for people with speech disorders, the Maximum Likelihood Linear Regression (MLLR) technique was used [20]. The language models for the baseline, the GA, and the Metamodels, consisted of WB and PB bigrams estimated from all the 740 sentences in the database.

From each dysarthric speaker, 18 randomly selected sentences were used for speaker adaptation and phoneme confusion-matrix estimation. Then, a different set of 18 sentences was selected for fitness evaluation of the GA (control). Finally, the resulting Metamodels of the GA optimization were tested with the remaining 38 sentences.

5.2 Convergence of the GA

Figure 7 shows the mean graph of fitness convergence across 20 executions of the GA and all dysarthric speakers. The Initial Metamodels are those built with the single confusion-matrix approach, the Baseline HMM is the performance of the adapted ASR system, and the Dyn-GA Metamodels are those built with the dynamic confusion-matrix approach optimized with the GA. As it is observed, the Initial Metamodels perform better than the baseline, something that is consistent with the results obtained in [3] and [12]. The Dyn-GA Metamodels increase their performance as the GA iterates, achieving a stable convergence after 10 iterations. It's important to mention that all of the Dyn-GA Metamodels obtained on each of the 20 executions of the GA achieved the same results on the test set. Hence, the results presented in Table 1 were obtained with a single execution of the GA.

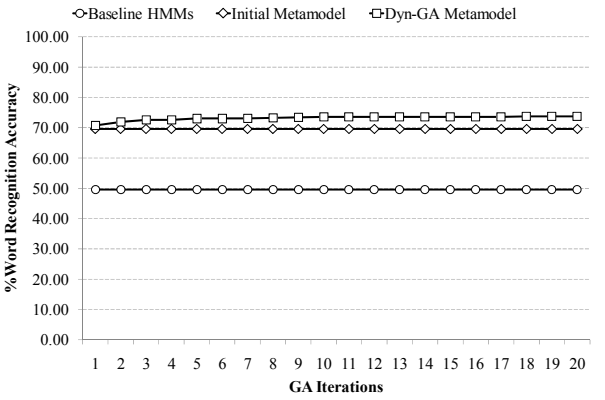

Fig. 7. Convergence of the GA

5.3 Results on Test Data

The mean recognition results across all dysarthric speakers on the test set (380 sentences) are presented in Table 1. A gain of 3.3% (4.5% absolute over the Initial Metamodels performance) is achieved with the Metamodels built with the proposed appoach. This gain was statistically significant as measured by the matched-pairs test described in [5] obtaining a p-value < 0.05.

Table 1. Mean %WAcc across all speakers on the test set

System	%Word Accuracy
Adapted Baseline	54.702
Initial Metamodels	73.406
Dyn-GA Metamodels	76.724

6 Conclusions and Future Work

In this paper we present an approach to obtain phoneme confusion-matrix estimates to improve the performance of Metamodels, a technique for dysarthric ASR. The approach consists of the dynamic estimation of confusion-matrices across different language model restrictions. A single confusion matrix then is estimated from the weighted confusion matrices previously obtained. The weights and other parameters of the Metamodels were optimized by means of a Genetic Algorithm (GA).

As presented in Section 5, with the proposed approach statistically significant gains were achieved over a standard baseline system and Metamodels when tested with a well known dysarthric speech database.

Future work will be aimed at (1) evaluating the performance of the approach with larger language models and other speech databases (TED, WSJ); (2) improving the GA with other operators for reproduction and selection; (3) to apply

the GA on other HMM parameters as the number of states and the observation probabilities associated to each HMM state; and (4) to explore the use of the dynamic phoneme confusion-matrices for speaker identification purposes.

References

1. Bunnel, H.T., Polikoff, J.B., Menéndez-Pidal, X., Peters, S.M., Leonzio, J.E.: The Nemours Database of Dysarthric Speech. In: Proc. International Conference on Spoken Language Processing (1996)
2. Caballero, S.O.: Structure Optimization of Metamodels to Improve Speech Recognition Accuracy. In: Proc. of the International Conference on Electronics Communications and Computers (CONIELECOMP 2011), pp. 125–130 (2011)
3. Caballero, S.O., Cox, S.J.: Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. EURASIP J. Adv. Signal Processing, 1–14 (2009)
4. Caballero, S.O., Cox, S.J.: On the Estimation and the Use of Confusion-Matrices for Improving ASR Accuracy. In: Proc. of Interspeech 2009 (2009)
5. Gillick, L., Cox, S.J.: Some statistical issues in the comparison of speech recognition algorithms. In: Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, pp. 532–535 (1989)
6. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Co. (1989)
7. Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M.S., Parker, M.: Automatic speech recognition with sparse training data for dysarthric speakers. In: Proc. European Conf. on Speech Communication Technology, pp. 1189–1192 (2003)
8. Hamidi, F., Baljko, M., Livingston, N., Spalteholz, L.: CanSpeak: A Customizable Speech Interface for People with Dysarthric Speech. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part 1. LNCS, vol. 6179, pp. 605–612. Springer, Heidelberg (2010)
9. Hasegawa-Johnson, M., Gunderson, J., Perlman, A., Huang, T.: HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (2006)
10. Kent, R.D., Wang, Y.T., Duffy, J.R., Thomas, J.E.: Dysarthria associated with traumatic brain injury: speaking rate and emphatic stress. Journal of Communication Disorders 38, 231–260 (2005)
11. Levit, M., Alshawi, H., Gorin, A., Nöth, E.: Context-Sensitive Evaluation and Correction of Phone Recognition Output. In: Proc. European Conference on Speech Communication and Technology, ISCA (2003)
12. Matsumasa, H., Takiguchi, T., Ariki, Y., Li, I.-C., Nakabayash, T.: Integration of Metamodel and Acoustic Model for Dysarthric Speech Recognition. Journal of Multimedia - JMM 4(4), 254–261 (2009)
13. Niu, X., Santen, J.P.: A formant-trajectory model and its usage in comparing coarticulatory effects in dysarthric and normal speech. In: Proc. Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), pp. 233–236 (2003)
14. Polur, P.D., Miller, G.E.: Effect of high frequency spectral components in computer recognition of dysarthric speech based on a Mel-cepstral stochastic model. Journal of Rehabilitation Research and Development 42(3), 363–372 (2005)
15. Raghavendra, P., Rosengren, E., Hunnicutt, S.: An investigation of different degrees of dysarthric speech as input to speaker adaptive and speaker dependent recognition systems. Aug. & Alt. Communication 17, 265–275 (2001)

16. Rosen, K., Yampolsky, S.: Automatic speech recognition and a review of its functioning with dysarthric speech. Aug. & Alt. Communication 16, 48–60 (2000)
17. Seong, W.K., Park, J.H., Kim, H.K.: Dysarthric Speech Recognition Error Correction Using Weighted Finite State Transducers Based on Context–Dependent Pronunciation Variation. In: Miesenberger, K., Karshmer, A., Penaz, P., Zagler, W. (eds.) ICCHP 2012, Part II. LNCS, vol. 7383, pp. 475–482. Springer, Heidelberg (2012)
18. Strik, H., Sanders, E., Ruiter, M., Beijer, L.: Automatic recognition of dutch dysarthric speech: a pilot study. In: ICSLP, pp. 661–664 (2002)
19. Wu, C.-H., Su, H.-Y., Shen, H.-P.: Articulation-Disordered Speech Recognition Using Speaker-Adaptive Acoustic Models and Personalized Articulation Patterns. ACM Transactions on Asian Language Information Processing (TALIP) 10(2) (2011)
20. Young, S., Woodland, P.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department (2006)