Applied Computing and Informatics (2011) 9, 1–10

King Saud University
Applied Computing and Informatics
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Adaptation of foreign accented speakers in native Arabic ASR systems

Sid-Ahmed Selouani [a,*], **Yousef Ajami Alotaibi** [b]

[a] *LARIHS Lab., Université de Moncton, Campus de Shippagan, Canada*
[b] *Computer Engineering Department, King Saud University, Saudi Arabia*

Received 1 February 2009; accepted 1 March 2010
Available online 16 December 2010

KEYWORDS

Arabic language;
Adaptation;
Foreign accents;
MLLR;
MAP;
HMM;
WestPoint;
LDC;
Native

Abstract This paper addresses the adaptation of Arabic speech recognition (ASR) systems to foreign accented speakers. This adaptation is accomplished by using the adaptation techniques; namely, the Maximum Likelihood Linear Regression (MLLR), the Maximum a posteriori (MAP), and the combination of MLLR and MAP. The LDC-WestPoint Modern Standard Arabic (MSA) corpus and HTK toolkit were used in implementing all experiments. The systems were evaluated using both word and phoneme levels. Results show that unique MSA Arabic Phonemes such as pharyngeal and emphatic consonants, which are difficult to pronounce for non-native speakers, benefit from the adaptation process using MLLR and MAP combination. An overall improvement of 7.37% has been obtained. This opens the eyes in benefiting from adaptation techniques in overcoming the difficulties of pronouncing nonnative language phonemes.

* Corresponding author.
E-mail addresses: selouani@umcs.ca (S.-A. Selouani), yaalotaibi@ksu.edu.sa (Y.A. Alotaibi).

Peer review under responsibility of King Saud University.
doi:10.1016/j.aci.2010.03.001

Production and hosting by Elsevier

1. Introduction

Characteristics of a second, non-native language are largely influenced by the first (native) language. As a result, the performance of automatic speech recognition systems, usually trained by native speakers, often degrades when they are used by non-native speakers. This is mainly due to both acoustic and phonological differences between accents (Hang et al., 2000; Zheng et al., 2005). These differences are not only due to different phoneme inventories of the languages, but even for the same phoneme, non-native and native speakers pronounce different sounds. The modeling of separate accents remains difficult and inaccurate due to the large number of non-native accents and to the insufficiency of non-native speech data available for training. It is the reason why many studies propose to adapt native phoneme models to accented phoneme models using first language data. Numerous studies have been carried out to improve the automatic recognition of speech uttered by non-native speakers.

Fakotakis (2004) worked on the adaptation of standard Greek speech recognition systems to work with Cypriot dialect by using HTK toolkit (Young, 2006), MLLR, MAP, and combined MLLR and MAP techniques (Lee and Gauvain, 1993; Leggeter and Woodland, 1995). He considered Cypriot Greek as a variation dialect of standard Greek with the same set of phonemes. He used utterances read from isolated digits, digit strings, application words, dates, and dictionary assistance names. In this study 500 native Greek speakers were involved in the training, while 450 speakers were used to adapt to the system, and 50 speakers to test it. When the system was trained by pure Cypriot Greek, the performance degraded due to inadequate training data. The best accuracy improvement was encountered with digits strings database and the combined MLLR and MAP technique. This improvement was 2.1%.

Bartkova and Jouvet (2004) proposed multiple models for improved speech recognition of non-native French speakers. They addressed the problem of foreign accent by using acoustic models of the target language phonemes (French phonemes in their case) adapted with speech data from three other languages: English, German, and Spanish. Their results obtained for 11 language groups of speakers showed that error rate can be significantly reduced when standard acoustic models of phonemes are adapted using speech data from other languages. In their outputs, the highest error rate reduction of 40% was obtained on English native speakers. They improved the recognition performance on almost all language groups, even though only three foreign languages were available in their study for acoustic model adaptation.

Hui et al. (2000) worked on principal mixture speaker adaptation for improved continuous speech recognition. They introduced a method that reduced HMM complexity by choosing only the principle mixtures corresponding to particular speaker's characteristics. This method improved both recognition accuracy (by

31%) and recognition speed (by 30%) when compared to full mixture speaker adaptation models.

The research on Arabic language mainly focuses on Modern standard Arabic, which is used throughout the media, courtrooms and academic institutions of the Arabic countries. Previous work on developing ASR was dedicated to dialectal and colloquial Arabic within the 1997 NIST benchmark evaluations, and more recently on the recognition of conversational and dialectal speech, as it is reported in Kirchhoff et al. (2003). Moreover, and compared to other languages, the Arabic language benefits from very limited number of research efforts. The goal of this paper is to investigate how adaptation techniques could improve a trained recognition system to be used by non-native Arabic speakers to get minimum amount of degradation in system accuracy. This adaptation is accomplished by using the adaptation techniques; namely, MLLR, MAP, and combination of MLLR and MAP. The original recognition system was designed for and trained by native Arabic speakers. Before adaptation, the system was tested by non-native Arabic speakers and the performance was considered for the sake comparisons with those of the adapted systems. We have four adaptation lists and three adaptation techniques; hence we have 12 adapted systems. Each system is evaluated at both word level and phoneme level.

The organization of this paper is as follows. The second section gives a basic background on Arabic language. In the third section, adaptation methods are briefly presented. Then, the fourth section presents the experimental framework, and the fifth section proceeds with a discussion of the obtained results. Finally, the sixth section concludes and indicates the perspective of this work.

2. Basic Arabic language background

Arabic is a Semitic language, and it is one of the oldest languages in the world today. It is the fifth widely used language nowadays (Al-Zabibi, 1990). The Arabic language has many differences when compared to European languages such as English. Some of the other differences are Arabic unique phonemes, phonetic features, and complicated morphological structures. A major difference lies in the Arabic text, where it is written with the absence of any information that leads to short vowels, geminate, and pharyngealization. This might lead to many identical-looking forms in a large variety of contexts, which decreases predictability in correct word pronunciation, sentence meaning, and language model rules. Hence, the determination of accurate language model from texts becomes very difficult when the type and position of short vowels, for example, are unknown (United Nations, 2003; Selouani and Caelen, 1998). Modern Standard Arabic (MSA) has 34 basic phonemes, of which six are vowels and 28 are consonants. The Arabic language has fewer vowels than English. It has three long and three short vowels, while American English has at least 12 vowels. Permissible syllables in the Arabic language include the following: CV, CVC, and CVCC, where V indicates a (long

or short) vowel, while C indicates a consonant. Arabic utterances can only start with a consonant.

Arabic is characterized by the presence of emphatic and pharyngeal phonemes. There are a total of five pharyngeal phonemes; among which two are fricatives: /H/ and /C/. These phonemes are characterized by the constriction formed between the tongue and the lower pharynx in addition to the rising of the larynx. There are three uvular pharyngeal phonemes, /x/, /G/, and /q/ characterized by a constriction formed between the tongue and the upper pharynx for /x/ and /G/ and a complete closure for /q/ at the same level. On the other hand, there are four emphatic phonemes: /S/, /D/, /T/, and /Z/. These phonemes are emphatic versions of the oral dental consonants /s/, /d/, /t/, and /TH/.

3. System adaptation methods

The widely-used adaptation technique is MLLR (Leggeter and Woodland, 1995). It is a parameter transformation technique that has proven successful while using a small amount of adaptation data. 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 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 MLLR transformation matrix used to give a new estimate of the adapted mean is stated as:

$$\tilde{\mu} = W\mu', \tag{1}$$

where W is the $n \times n(n+1)$ transformation matrix (where n is the dimensionality of the data) and μ' is the extended mean vector defined as follows:

$$\mu' = [w\,\mu_1\,\mu_2\,\ldots\,\mu_n]^T, \tag{2}$$

where w represents a bias offset whose value here is fixed at 1. Hence W can be decomposed into:

$$W = [b_c\,A_c]$$

where A_c represents an $n \times n$ regression matrix and b_c is an additive bias vector associated with the broad class c. The adapted kth mean vector for each state i can be written as follows:

$$\tilde{\mu}_{ik} = A_c\mu_{ik} + b_c. \tag{3}$$

The system adaptation can be accomplished using Maximum a posteriori (MAP) technique (Lee and Gauvain, 1993). For MAP adaptation, the re-estimation formula for Gaussian mean is a weighted sum of the prior mean with the maximum likelihood mean estimate. It is formulated as:

$$\hat{\mu}_{ik} = \frac{\tau_{ik}\mu_{ik} + \sum_{t=1}^{T}\xi_t(i,k)x_t}{\tau_{ik} + \sum_{t=1}^{T}\xi_t(i,k)}, \tag{4}$$

where τ_{ik} is the weighting parameter for the kth Gaussian component in the state i. $\xi_t(i,k)$ is the occupation likelihood of the observed adaptation data x_t.

One of the drawbacks of MAP adaptation is that it requires more adaptation data to be effective compared to MLLR. When MLLR is combined with MAP we can benefit from both of the techniques. Theoretically, the combination offers compact transformations for rapid adaptation when only limited amount of data is available, thanks to MLLR, and the asymptotical efficacy of MAP adaptation when the amount of data increases. There are many ways to combine MLLR and MAP. We choose to use the MLLR transformed means as the priors for MAP adaptation. Hence, the adapted means can be written as:

$$\hat{\mu}_{ik} = \frac{\tau_{ik}\tilde{\mu}_{ik} + \sum_{t=1}^{T}\xi_t(i,k)x_t}{\tau_{ik} + \sum_{t=1}^{T}\xi_t(i,k)}. \tag{5}$$

The principal difficulty in MAP adaptation is to determine the mixing parameters. As it is commonly used, we chose a single mixing parameter for each model that we built, i.e., $\tau_{\mathrm{ik}} = \tau$.

4. Experimental setup

The WestPoint Arabic Corpus, provided by LDC (2002), is used in our experiments. It consists of collections of four main Arabic scripts. First is Collection Script 1, which contains 155 sentences, used by all 74 native Arabic speakers. Script 1 has a total of 1152 tokens and 724 types. Second is Collection Script 2, which contains 40 sentences, used by 23 of the non-native speakers. Script 2 has a total of 150 tokens and 124 types. Third is Collection Script 3, which contains 41 sentences, used by 4 of the non-native speakers. It has a total of 138 tokens and 84 types. Finally, there is Collection Script 4, which contains 22 sentences, used by 9 of the non-native speakers, all of them are third year Arabic speakers. It has a total of 72 tokens and 59 types. The total number of distinct words is 1131 Arabic words. All scripts were written with MSA as the target language and were diacritized. In the LDC-Westpoint database, the amount of data provided by the Arabic native speakers is significantly larger than that of the data provided by the Arabic non-native speakers. Note that the WestPoint corpus has three phonemes more than the number of MSA phonemes mentioned in the linguistic literature (Kirchhoff et al., 2003; Ouni et al., 2005). These phonemes are: /g/ ''voiced velar stop'', /aw/ ''back upgliding diphthong'', and /ey/ ''upper mid front vowel''. In fact, the phoneme /g/ does not exist in MSA at all, but we think that the WestPoint corpus used it because some native and non-native speakers are using it popularly in some MSA words. We can confirm

this fact by hearing some WestPoint audio files. On the other hand, the extra vowel and diphthong were used because of variations in pronunciations of speakers influenced by English and other Latin languages. This type of phoneme exists in these languages but not in MSA. For our study, we finally decided to stick with WestPoint phonemes and transcriptions without any modification. We considered this to make it more easy and logical to compare these results with other researches results which used the same corpus. Using the same settings and variables values will give more correct and comparable outcomes.

4.1. Data

From the WestPoint corpus we selected four different and disjointed lists; all have been chosen randomly from non-native Arabic speakers only. The first list called AD100, contains 100 utterances; the second list called AD150, contains 150 utterances; the third list called AD200, contains 200 utterances; the last list called AD250, contains 250 utterances. The four lists are chosen randomly from all available scripts, speakers, and genders. The designed lists were used to adapt a native Arabic speaker based system to deal with non-native Arabic speakers. For this purpose, three different adaptation techniques were used: MLLR, MAP and a combination of MLLR and MAP. The performance is analyzed both at the word level (by incorporating a language model) and at the phoneme level. The phoneme level permits us to investigate the improvement (if any) of system accuracy on individual phonemes; hence giving us the chance to analyze the weakness of non-native Arabic speakers' pronunciation as a phoneme-wise way of analysis. Based on the above explanations, we refer to our experiments as AD100/MLLR, AD100/MAP, AD100/MLLRMAP, etc.

4.2. Recognition platform and parameters

The Hidden Markov Model Toolkit (HTK) (Young et al., 2006) is used for designing and testing the speech recognition systems throughout all experiments. The baseline system was initially designed as a phoneme level recognizer with three active states, continuous, left-to-right, no skip HMM models. The system was designed by considering all 37 MSA phones as given by the LDC catalog. Since most of the words consisted of more than two phonemes, context-dependent triphone models were created from monophone models. The training phase consists of re-estimating HMM models by using Baum-Welch algorithm after aligning and tying the models by using the decision tree method. Phoneme-based models are good at capturing phonetic details. Also context-dependent phoneme models can be used to characterize formant transition information, which is very important for the discrimination of confusable speech units.

The parameters of the system consist of a 22 kHz sampling rate with a 16 bit sample resolution, a 25 ms Hamming window duration with a step size of

10 ms, MFCC coefficients with 22 as the length of cepstral leftering and 26 filter bank channels of which 13 were the number of MFCC coefficients, and of which 0.95 were the pre-emphasis coefficients.

5. Results and discussion

All native Arabic speakers' data provided by the LDC WestPoint corpus were used for training the original recognition system. After that, all non-native Arabic speakers' data provided by the same corpus was used for testing the system. As a result of that test, the accuracy (correctness) of the system was 89.02% and 93.19% for word level and phoneme level, respectively. This performance is relatively low compared to the same system but with testing data taken from native Arabic speakers where the percentage for word level and the percentage for phoneme level have been, respectively, obtained. Fig. 1 shows the system performance for the four adaptation lists and for the three adaptation techniques. This performance is compared to the accuracy of the system prior to any kind of adaptation.

As it can be inferred from the results (cf. Fig. 1), the improvement of system performance increases when the size of the adaptation list is increased. This performance is improved rapidly to reach its best at 96.39% which represents a 7.37% improvement (at word level) in comparison with the original system. This result is obtained by the adaptation list AD250 and the adaptation combining MLLR and MAP techniques (i.e., experiment AD250/MLLRMAP). We noticed that in AD150 and AD250, the combined MLLR and MAP adaptation techniques gave better performance compared to others.

The improvements in accuracy for the different experiments are depicted in Table 1. We notice that there is no fixed rule governing the comparisons of

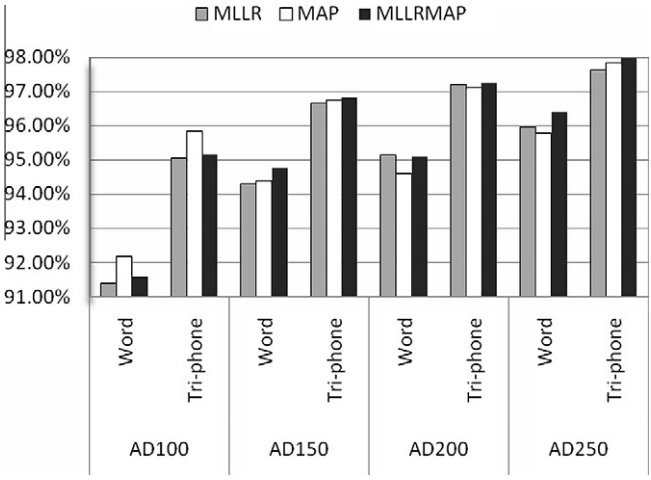

Figure 1 System accuracies for MLLR, MAP and MLLR/MAP techniques with different sizes of adaptation data.

Table 1 Accuracy improvement using adaptation techniques with different sizes of adaptation data.

Adaptation list	Level	MLLR (%)	MAP (%)	MLLRMAP (%)
AD100	Word	2.37	3.16	2.56
	Tri-phone	1.88	2.65	1.97
AD150	Word	5.27	5.37	5.74
	Tri-phone	3.47	3.57	3.64
AD200	Word	6.12	5.58	6.08
	Tri-phone	4.02	3.92	4.06
AB250	Word	6.95	6.77	7.37
	Tri-phone	4.45	4.64	4.80

MLLR, MAP, and their combination. In some experiments, MLLR gave improvement better that that of MAP. In other experiments MAP gave better accuracy improvement. The combined MLLR and MAP techniques sometimes gave less improvement compared to either MLLR or MAP. For instance, experiment AD100/MLLRMAP gave 1.97% as accuracy improvement but AD100/MAP gave better performance with a 2.65% improvement. We believe that this is due to the random choice of sentences used in adaptation. In some cases, more relevant and specific Arabic phonemes are included in the adaptation data, while in other cases, the adaptation set contains less of these phonemes. As a general observation, we noticed that MAP gave better accuracy improvements compared to MLLR, and MLLRMAP gave generally better accuracy improvements compared to MAP and MLLR.

By investigating the system performances for individual phonemes we can notice that the phonemes /H/, /TH/, /g/, /q/, and /z/ gained more improvement in their performances for all experiments. Table 2 shows the increases in performance for these phonemes for all conducted experiments. Except for phoneme /z/, these phonemes are Arabic phonemes that cannot be found in English. It is a very

Table 2 Accuracy improvement after adaptation for 4 Arabic specific phonemes.

Adapt list	Adapt technique	Improvement in accuracies (%)				
		/H/	/TH/	/g/	/q/	/z/
AD100	MLLR	6.9	4.5	11.1	6	8.8
	MAP	6.1	1	10.7	5.6	8.8
	MLLRMAP	6.9	5.4	6.2	6.9	10.3
AD150	MLLR	6.8	5.8	18.4	7.5	8.8
	MAP	6.8	4.5	3.8	7.7	9.1
	MLLRMAP	8.2	5.4	13.6	7.3	10.3
AD200	MLLR	10.4	5.4	8.6	8	10.7
	MAP	8.2	3.6	8.4	8	10.3
	MLLRMAP	10.6	5.8	6.2	8.6	11
AD250	MLLR	9.6	6.7	16	7.3	11.7
	MAP	11.2	4.9	13.6	9.1	12.1
	MLLRMAP	10.4	6.7	16	7.8	11.4

encouraging result since the adaptation process permits the enhancement of the performance of phonemes that are hard for non-native Arabic speakers to pronounce, especially /H/ which is a fricative unvoiced non-emphatic pharyngeal sound. Thus, in the non-adapted system these sounds and other particular Arabic phonemes produced errors due to the phonological and acoustical changes induced by the pronunciation of non-native speakers. Consequently, these results confirm that by the means of the adaptation process, the performance of automatic recognition of Arabic foreign-accented speech can be significantly improved.

6. Conclusion and future work

An automatic Arabic speech recognition system was trained by using native Arabic speakers' speech data provided by the LDC WestPoint corpus for MSA Arabic. After that, the system was adapted to non-native Arabic speakers' speech data provided by the same corpus. The adaptation techniques were MLLR, MAP, and a combination of them. The accuracies of the non-adapted system were 89.02% for word level accuracy and 93.19% for phoneme level accuracy. The best system accuracy improvement was 7.37% and this was obtained in experiment AD250/MLLRMAP. Among others, the specific Arabic phonemes /H/, /TH/, /q/, and /H/ known as being hard to pronounce for a non-native Arabic speaker got better accuracy improvements in all experiments.

This work will be continued by investigating the performance of an evolutionary-based technique to give the Arabic speech recognition system an auto-adaptation capability in the context of more foreign accents. Also this work will be expanded to use a better in quality and bigger in size speech corpora. In Addition to this, we are planning to check the effects of other languages (in addition to English) as a first language of speakers on improving the performance of Arabic ASR systems. Also we can do it for the opposite way in improving English ASR systems in case of using them by Arabic native speakers whom can pronounce unique (compared to Arabic) English phonemes such as /p/, /v/, and /g/.

References

Al-Zabibi, M., 1990. An Acoustic-Phonetic Approach in Automatic Arabic Speech Recognition. The British Library in Association with UMI.

Bartkova, K., Jouvet, D. 2004. Multiple models for improved speech recognition for non-native speakers. In: SPECOM'2004: 9th Conference of Speech and Computer, St. Petersburg, Russia.

Fakotakis, N. 2004. Cypriot speech database: data collection and greek to cypriot dialect adaptation. In: 4th International on Language Resources and Evaluation, May 24–30.

Hang, C., Change, E., Zhou, J., Lee, K.-F. 2000. Accent modeling based in on pronunciation dictionary adaptation for large vocabulary mandarin speech recognition. In: International Conference on Speech and Language Processing (ICSLP), Beijing, pp. 818–821.

Hui, Y., Fung, P., Taiyi, H. 2000. Principal mixture speaker adaptation for improved continuous speech recognition. In: International Conference on Speech and Language Processing (ICSLP), Beijing.

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Gang, J., Feng, H., Henderson, J., Daben, L., Noamany, M., Schone, P., Schwartz, R., Vergyri, D. 2003. Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop. In: Proceedings of ICASSP 2003, April, vol. 1, pp.344–347.

LDC. 2002. Linguistic Data Consortium (LDC) Catalog Number LDC2002S02. <http://www.ldc.upenn.edu/>.

Lee, C.-H., Gauvain, J.-L. 1993. Speaker adaptation based on MAP estimation of HMM parameters. In: Proceedings ICASSP, Minneapolis, Minnesota, pp. 558–561.

Leggeter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. Computer Speech and Language 9, 171–185.

Ouni, S., Cohen, M., Massaro, W., 2005. Training Baldi to be multilingual: a case study for an Arabic Badr. Speech Communication 45, 115–137.

Selouani, S., Caelen, J. 1998. Arabic phonetic features recognition using modular connectionist architectures. In: Interactive Voice Technology for Communication, IVTTA '98, 98 IEEE 4th Workshop 29–30 Sept. 98, Torino, Italy, pp. 155–160.

United Nations. 2003. Economic and Social Commission for Western Asia–United Nations Report, Harmonization of ICT Standards Related to Arabic Language Use in Information Society Applications, United Nations Report, New York.

Young (2006). "Using POMDPs for Dialog Management." IEEE/ACL Workshop on Spoken Language Technology (SLT 2006), Aruba.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. 2006. HTK Version. 3.4, Cambridge University Engineering Department. <http://htk.eng.cam.ac.uk/prot-doc/ktkbook.pdf>.

Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., Yoon, S. 2005. Accent detection and speech recognition for shanghai-accented mandarin. In: 9th European Conference on Speech Communication and Technology (Eurospeech), Lisboa, Portugal.