Recognizing Arabic Speakers with English Phones

Andreas Stolcke *Sachin Kajarekar*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
{stolcke,sachin}@speech.sri.com

Abstract

We investigate the question of whether phone recognition models trained on large English databases can be used for speaker recognition in another language. Such a cross-language use of recognition models is an attractive option when a speaker recognition system is to be ported to a new language without the necessary data resources, while retaining some of the advantages of phone modeling and ASR-based feature extraction. We compare the performance of such systems to a baseline cepstral GMM system (which is inherently language independent), and to a phone-recognition-based system trained exclusively on Arabic data. Our results indicate that cross-language models are highly competitive, and, at least in our case, have a performance advantage over within-language training and the language-independent baseline. We also examine the effect of coverage of colloquial Arabic dialects in the training data.

1. Introduction

Recent years have seen advances in speaker recognition through the use of phone recognition models. One popular approach is to decode ("tokenize") the speech sample by unconstrained phone recognition, and then to model the phone sequences [1, 2, 3]. Another approach uses maximum likelihood linear regression (MLLR) speaker adaptation transforms computed against a phone-loop model, followed by support vector machine (SVM) classification of the transform coefficient [4, 5]. Both types of model achieve performance comparable or even exceeding that of standard cepstral Gaussian mixture models (GMMs), but they are based on language-specific phone recognition models. In all cases one finds significant gains from combining the phone-based systems with the baseline GMM systems. This makes sense since the phone-based systems capture longer-term features of the speech signal (in the case of phone sequences) or represent the observations in a very different feature space, thereby providing complementary information to the speaker recognition system.

The standard approach for porting such speaker recognition systems to a new language would require redesigning the phone set and training phone models on target language data. However, such an approach would be quite involved, and might not even be feasible if available transcribed speech data is limited. Fortunately, unconstrained phone recognition is fast and flexible enough to be run on mismatched languages; the recognizer will simply choose the phones providing the best acoustic match to the foreign language. As a result, phone recognition will generate an (imperfect) representation of the new language in terms of the old phone set, a principle that has long been explored for language recognition [6].

Our goal, then, is to leverage existing English phone recognition models to extract features in a non-English language, specifically for Arabic. In this study we investigate how well such a cross-language approach works for speaker recognition, and how it compares to native retraining of the phone models. We also compare phone-recognition-based systems to a cepstral GMM baseline, which is by its nature independent of language.

2. Data

We used data from three Arabic dialects to train the background (speaker-independent) model in the GMM and MLLR-SVM systems. (Note that for background training data, broadcast speech is included, to increase the dataset size.) All experiments were conducted using an 8 kHz sampling rate.

- Modern Standard Arabic (MSA) is the dialect used in formal communication. The data was collected from radio newscasts from various radio stations in the Arabic-speaking world by the Foreign Broadcast Information Service (FBIS). It contains 145 recordings with an average length of 15 minutes.

- Levantine Arabic (LVA) is a group of dialects spoken in the Levant (Syria, Palestine/Israel, western Jordan and Lebanon). The data includes 544 telephone conversations with an average length of 5 minutes.

- Egyptian Arabic (EGA) is widely understood in Egypt and many other Arab countries. This data contains 120 telephone conversations collected in

the Linguistic Data Consortium's (LDC) Call-Friend setup, with an average length of 5 minutes.

For the study exploring variation in background data sets in Section 4.3, we also used data for the following additional dialects:

- Iraqi Arabic (IA), comprising 478 conversation sides from an LDC telephone collection. The average duration of a conversation is about 6 minutes.
- Gulf Arabic (GA), comprising 526 conversation sides from an LDC telephone collection, with average duration of 5.7 minutes per call.

For testing we used all Arabic-language conversations (of unknown dialect) contained in the NIST SRE-04 and SRE-05 evaluation corpora, a subset of the LDC Mixer corpus [7]. This dataset contains speech from 43 speakers with an average of 5 conversations per speaker, 594 target trials, and 5940 impostor trials. Note that the available data only allowed for trials using a single training conversation side per target speaker.

3. Systems

3.1. Cepstral GMM

A GMM system was used to model Mel-cepstral features, including deltas and double-deltas. The system was based on the GMM-UBM paradigm [8], where a speaker model is adapted from a universal background model (UBM). Maximum a posteriori (MAP) adaptation was used to derive a speaker model from the UBM. The GMM had 2048 Gaussian components. The cepstral GMM system includes gender/handset normalization and utterance-level mean and variance normalization. Two background models were used, one trained with English data from the Switchboard (landline and cellular) and Fisher databases, and another with the Arabic data described above.

3.2. Phone-loop MLLR SVM

The second model is a maximum likelihood linear regression MLLR-SVM [4] system. It estimates adaptation transforms for each speaker, using a phone-loop speech model with three regression classes, for nonspeech, obstruents, and nonobstruents (the nonspeech transform is not used). Such a system models speaker-specific translations of the Gaussian means of phone recognition models, and does not require running a word recognition system. We used an English phone recognition system (with an English phone set and trained on English Switchboard telephone data), based on a 39-dimensional feature vector derived from Mel-cepstra, voicing features, deltas, and double-deltas. The phone models were trained on English conversational telephone speech (CTS) databases, namely, the Switchboard-I corpus and a small subset of the Switchboard-II cellular corpus.

Since the phone models were gender dependent, we computed two sets of transforms, one for each gender model. Transform coefficients from the two models and the two phone classes are concatenated to form a $2 \times 2 \times 39 \times 40$-dimensional feature vector. Each feature dimension is rank normalized to the unit interval using the Arabic background data as the reference distribution. Finally, a linear inner-product kernel SVM is trained for each target speaker using the feature vectors from the background training set as negative examples, and the target speaker training data as positive examples. The speaker verification score is the signed distance of the test sample vector from the decision hyperplane.

3.3. Phone N-gram SVM

This is an SVM version of the widely used phone sequence modeling, based on phone lattices rather than 1-best recognition output [3]. An open-loop phone recognizer (trained on English Switchboard data) is run on each conversation side, generating lattices. We then extract expected frequencies for unigrams, bigrams, and trigrams, (i.e., N-grams are weighted according to their posterior probability of occurrence in the lattice). The 14k most frequent N-grams (extracted from the background data) are retained, giving the dimensionality of the feature vector. The N-gram frequencies are then scaled by the inverse square roots of the overall N-gram probabilities. When combined with a linear SVM kernel, this gives the log likelihood ratio kernel of [2].

4. Experiments and Results

4.1. Results with English phone models

Our first experiments establish the GMM baseline result and compare the phone-based systems to that baseline. Note again that the phone recognition models were trained on English data, but then applied to Arabic background, target speaker and test data. For the GMM we trained two versions of the background model: one using English CTS data, and one using the Arabic background set. Table 1 summarizes the results, reported in terms of equal error rate (EER).

We can characterize the results as follows. Using Arabic versus English data for UBM training does make a difference, albeit a surprisingly small one (11.5% relative). This can be rationalized by the unstructured nature of the GMM, as well as the fact that the English background set, while mismatched, is much larger than the Arabic one. The phone-recognition based systems perform roughly on par with the cepstral GMM. The phone N-gram SVM is somewhat worse (22.2% relative), while the MLLR-SVM is slightly better (7.5% relative).

Also shown in Table 1 are two simple combination re-

Table 1: Individual and combined system results on Arabic test data

System	Bkg. data	%EER
Cepstral GMM	English	10.27
Cepstral GMM	Arabic	9.09
Phone N-gram SVM	Arabic	11.11
Phone-loop MLLR SVM	Arabic	8.41
Combined Systems		
Phone N-gram + MLLR SVM	Arabic	7.74
Same + GMM	Arabic	7.45

Table 2: Comparison of speaker verification on Arabic and English (SRE-06) data

System	Test data language	
	Arabic % EER	English % EER
Cepstral GMM	9.09	7.16
Phone N-gram SVM	11.11	12.75
Phone-loop MLLR SVM	8.41	7.91

Table 3: Comparison of phone models trained on Arabic and English

System	Phone models trained on	
	Arabic % EER	English % EER
Phone N-gram SVM	19.70	11.11
Phone-loop MLLR SVM (m+f)	n/a	8.41
Phone-loop MLLR SVM (f only)	10.44	9.60

Table 4: Comparison of systems differing in choice of background data dialects

Background data	MLLR % EER	Phone N-gram % EER
ECA+LVA	8.42	11.45
ECA+LVA+MM	8.42	11.11
ECA+LVA+MM+IA	8.24	10.94
ECA+LVA+MM+IA+GA	8.92	10.94

sults, obtained by averaging two and three system scores with equal weight (the scarcity of data and comparable performance of all three systems favored this approach over a more elaborate trainable combiner, such as a neural network). The two phone-based systems combined give a 7.8% relative reduction in EER over the best single system (MLLR-SVM), while adding the cepstral GMM into the mix gives 11.4% relative reduction.

It is useful to compare the performance of all three systems to similar systems when tested on an English speaker recognition task. For this purpose we use the results obtained on the English-language subset (Common Condition, 1-side training) of the NIST SRE-05 dataset. Results are shown in Table 2. Remarkably, the pattern of results differs substantially across languages. Relative to the other systems, the phone N-gram SVM performs much better in Arabic, achieving an EER that is better than its English counterpart. The other systems have lower EERs in English, but the ordering is reversed, with the the cepstral GMM being somewhat better than the MLLR-SVM.

4.2. Results with Arabic phone models

Next we tested phone-based speaker recognition with phone models trained on Arabic data. This raised some issues since there was no additional CTS-like data available for training such models (training phone recognizer and speaker models on the same data would lead to severe bias in the models and mismatch with unseen training data). We decided to use Modern Standard Arabic models trained on broadcast data with a telephone (band-limited) front end configuration. While MSA is not necessarily a perfect match to the conversational data found in our training set, it should at least be closer to the target language than English.

Another issue is that the English models were gender dependent, giving us added leverage from combining male- and female-specific MLLR transforms as features. The MSA models we had available, by contrast, are gender independent, yielding only half the number of feature components. To compensate for this difference we also tested an English MLLR-SVM based on only one gender model.

Table 3 summarizes the results and compares to corresponding results obtained with the English phone models. We see that for both phone N-gram and MLLR SVMs, the English-trained models perform better as speaker feature extractors. This is true for MLLR SVM even after equating the feature dimensions by using only transforms based on female models.

4.3. Effect of background data

Finally, we wish to assess the effect of varying sources of Arabic dialectal data in the background (impostor) training set for SVM training. It is to be expected that results would improve the more different dialects are represented in the background data. To test this conjecture we tested a series of MLLR and phone N-gram SVM systems that incorporated an increasing variety of dialects. Only English phone models were used in these experiments. Results are given in Table 4.

The experiments show that, up to a point, results do

improve modestly as more and more dialects are covered in the background data. One exception to this pattern is the Gulf Arabic (GA) corpus, which degrades results for the MLLR-SVM system, and does not improve the phone N-gram SVM results. Experiments with other subsets of corpora including the GA data confirm this detrimental effect, and suggest that there may be some mismatch between this data source and the training data. A more detailed analysis of the data will be needed to pinpoint the reasons for this anomaly.

5. Conclusions and Future Work

We have studied the performance of phone-recognition-based speaker models on an Arabic speaker verification task. Both a phone N-gram SVM and a phone-loop MLLR SVM gave results comparable to or better than a standard cepstral SVM, with even better results when these models were combined with the GMM.

Remarkably, the English-trained phone models performed better on Arabic data than similar models trained on Modern Standard Arabic. This indicates that language mismatch in the phone models is less important than other factors, such as the amount of available data. It could also be that the English phone set, being relatively large (and larger than the Arabic phone set) gives enough acoustic resolution to be generally useful across languages. In any case, the cross-language use of English phone models for speaker recognition suggests itself as a viable strategy, especially where little matched-language training data is available.

An important issue for future work is a detailed investigation of how the amount of training data affects performance. For example, how would Arabic and English models compare if they had been trained on equivalent amounts of data? Another question is how matched and mismatched language data could be combined for the best effect.

Finally, the coverage of Arabic dialects merits further study. For example, it is possible that sampling the various dialectal data sources to better match the target population would improve results.

6. Acknowledgments

Thanks to Martin Graciarena for helping to establish the baseline system for our experiments as a result of his work on noise robustness in Arabic speaker recognition [9]. Thanks also to Dimitra Vergyri for help with the Arabic recognition system. This research was funded by award NMA401-02-9-2001 and a development contract with Sandia National Laboratories. The views herein are those of the authors and do not reflect the views of the funding agencies.

7. References

[1] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition", *in Proc. ICASSP*, vol. 1, pp. 149–152, Orlando, FL, May 2002.

[2] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines", in S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pp. 1377–1384, Cambridge, MA, 2004. MIT Press.

[3] A. O. Hatch, B. Peskin, and A. Stolcke, "Improved phonetic speaker recognition using lattice decoding", *in Proc. ICASSP*, vol. 1, pp. 169–172, Philadelphia, Mar. 2005.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition", *in Proc. Interspeech*, pp. 2425–2428, Lisbon, Sep. 2005.

[5] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987–1998, Sep. 2007.

[6] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling", *in Proc. ICASSP*, vol. 1, pp. 305–308, Adelaide, Australia, 1994.

[7] A. Martin, D. Miller, M. Przybocki, J. Campbell, and H. Nakasone, "Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004", *in Proceedings 4th International Conference on Language Resources and Evaluation*, pp. 587–590, Lisbon, May 2004.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 181–202, 2000.

[9] M. Graciarena, S. Kajarekar, A. Stolcke, and E. Shriberg, "Noise robust speaker identification for spontaneous Arabic speech", *in Proc. ICASSP*, vol. 4, pp. 245–248, Honolulu, Apr. 2007.