NOISE ROBUST SPEAKER IDENTIFICATION FOR SPONTANEOUS ARABIC SPEECH

Martin Graciarena[1], *Sachin Kajarekar*[1], *Andreas Stolcke*[1,2], *Elizabeth Shriberg*[1,2]

[1]SRI International, Menlo Park, CA, USA,
[2]International Computer Science Institute, Berkeley, CA, USA
www.speech.sri.com

ABSTRACT

Two important challenges for speaker recognition applications are noise robustness and portability to new languages. We present an approach that integrates multiple components and models for improved speaker identification in spontaneous Arabic speech in adverse acoustic conditions. We used two different acoustic speaker models: cepstral Gaussian mixture models (GMM) and maximum likelihood linear regression support vector machine (MLLR-SVM) models and a neural network combiner. The noise-robust components are Wiener filtering, speech-nonspeech segmentation, and frame selection. We present baselines and results on the Arabic portion of the NIST Mixer data, in clean conditions and with added noise at different signal-to-noise ratios. We used two realistic noises: babble and city traffic. In both noisy scenarios, we found significant equal error rate (EER) reductions over the no-compensation condition. The various noise robustness methods gave complementary gains for both acoustic models. Finally, the combiner provides a reduction in EER over the individual systems in noisy conditions.

***Index Terms*—** Speaker identification, Robustness, Arabic, cepstral GMM, MLLR-SVM

1. INTRODUCTION

Two increasingly important challenges for speaker recognition applications are noise robustness and portability to new languages. In this work, we focus on spontaneous Arabic speech data. Various methods have been proposed in order to reduce the influence of noise and achieve reasonable speaker identification accuracy in noisy acoustic scenarios. One of them is the Wiener filter, which aims at estimating a clean speech waveform from a noisy one. It has been successfully used in speaker identification systems [1]. Another technique is frame selection [2], whose goal is to only use frames in scoring which are less degraded by noise.

In this paper we integrate multiple components for improved speaker identification in adverse acoustic conditions. We first apply the Wiener filter to the unsegmented waveform. Next we extract speech-like segments from the unsegmented waveform. Our segmentation system uses SRI's Decipher speech recognition system and large vocabulary acoustic models. Finally, we do frame selection before scoring the speaker model. We select the frames that have an average energy above a certain threshold. In the experiments we added noise samples to clean speech at multiple signal-to-noise ratio (SNR) values. We used two realistic noises: babble and city traffic. We used state-of-the-art speaker identification systems and a combiner for the Arabic data. Scores were generated by a Gaussian mixture model (GMM) base system and a maximum likelihood linear regression and support vector machine (MLLR-SVM) system. Speaker models were trained with clean speech only. We also used a neural network score combiner.

All of the clean and noisy speech experiments were done using Arabic language data. We could find very few papers [3] on speaker identification in spoken Arabic.

We have found significant equal error rate (EER) reductions over the baseline (no compensation) condition and complementary gains using the proposed noise robust techniques in both types of noise. We also found the combiner to be quite robust to noise.

2. ARABIC 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.)

- 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 LDC's CallFriend setup, with an average length of 5 minutes.

All experiments were conducted using an 8 kHz sample rate. For testing we used all Arabic-language conversations (of unknown dialect) contained in the NIST SRE 04 and 05 [4] evaluation corpora, a subset of the LDC Mixer corpus. This dataset contains speech from 43 speakers with an average of 5 conversations per speaker, 594 target trials, and 5940 impostor trials.

1-4244-0728-1/07/$20.00 ©2007 IEEE

3. BASELINE SYSTEMS AND COMBINER

3.1. Cepstral Gaussian Mixture Model

A GMM system was used to model speaker-specific cepstral features. The system was based on the GMM-UBM model paradigm, 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 has 2048 Gaussian components, and is described in detail in [5]. The cepstral GMM system includes gender/handset normalization and utterance-level mean and variance normalization. Table 1 presents the results of the cepstral GMM system on the Arabic portion of the NIST Mixer data. 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. Maximum Likelihood Linear Regression Model

The second model is a maximum likelihood linear regression MLLR-SVM [6] system. It estimates adaptation transforms 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). The coefficients from the two speech adaptation transforms are concatenated into a single feature vector and modeled using support vector machines (SVMs). 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. Rank normalization on each feature dimension was used. Table 1 shows the results of the MLLR-SVM system on the Arabic Mixer data.

3.3. Score Combiner

The MLLR-SVM system is an acoustic model using cepstral features, but using a nonstandard representation of the acoustic observations. Therefore it may provide complementary information to the cepstral GMM. We used a neural network classifier from the LNKnet library [7] to combine the two systems at the score level. The neural network had two inputs, no hidden layer, and a single linear output activation unit. We split the testset in two halves for jackknifing purposes. The combiner was first trained using both systems' scores from the first half of the database and score estimates were generated for the second half. The procedure was repeated, but switching the training and test sets. Finally, we juxtaposed the estimated scores from both halves and computed the EER.

Table 1: EER Results without Added Noise on Arabic Mixer Data

	System	Background Data	% EER
1	GMM	English	10.27
2	GMM	Arabic	9.09
3	MLLR-SVM	Arabic	8.41
	Combiner (2+3)		8.42

Table 1 shows that there is a 10% improvement in the GMM system using Arabic background data compared to using English, even though the Arabic data set is much smaller. We can compare the 9.09% EER obtained on Arabic data to the 7.17% EER from the GMM system on the English SRE 05 test set [8], which has more and better-matched training data. The MLLR-SVM system is competitive with the best GMM system even though it uses English acoustic models for recognition and background modeling. The combiner does not produce a gain over the best system in clean conditions. However, it proves valuable in noisy conditions, as we will see next.

4. NOISY SPEECH PROCESSING TECHNIQUES

4.1. Wiener Filtering

The goal of the Wiener filter is to estimate a clean speech waveform from a noisy speech waveform. We used an implementation from the Qualcomm-ICSI-OGI Aurora system [9]. It first uses a neural-network-based voice activity detector to mark frames as speech or nonspeech. Next, a noise spectrum is estimated as the average spectrum from the nonspeech frames. Finally, this noise spectrum is used in the Wiener filtering of the noisy waveform. The Wiener filter was applied to the unsegmented waveform in order to take advantage of the long silence segments between speech segments for noise estimation.

4.2. Segmentation

We applied a speech-nonspeech segmenter to extract speech segments from the noisy speech waveform. This segmenter takes advantage of the cleaner signal produced by the Wiener filtering. The segmenter is from SRI's large-vocabulary telephone speech recognizer and was trained on 315 English telephone conversations from the Switchboard and CallHome corpora. It was not specifically tuned for Arabic or the noise conditions. Segmentation is performed by Viterbi-decoding each conversation side separately, using a speech/nonspeech hidden Markov model (HMM), followed by padding at the boundaries and merging of segments separated by short pauses.

4.3. Frame Selection

In GMM scoring only the frames with average frame energy above a certain threshold were used. In clean conditions it is desirable to discard silence frames. Therefore the energy threshold should be low in these conditions. In noisy conditions, however, we want to discard frames that are more likely to be degraded by noise. Here the threshold should be higher to eliminate noisy nonspeech frames frames. The actual energy threshold for each waveform is computed multiplying an energy percent (EPC) parameter (between zero and one) to the difference between maximum and minimum frame log energy values and adding the minimum log energy. We have found that the optimal EPC (i.e. the parameter for which the testset EER is the lowest) is dependent on both noise type and SNR.

5. EXPERIMENTS WITH NOISY ARABIC SPEECH

Here we show EER results in noisy conditions. We also test the proposed noise compensation techniques with both systems. Finally, we test how the combiner performs in noisy conditions. The noises used are babble noise from the Noisex-92 [10] database, and city traffic noise. They were digitally mixed with the clean speech waveforms at different SNRs. In all the experiments we used only speaker models trained on clean speech, since it is not possible to train speaker models in all noisy acoustic scenarios.

5.1. GMM System

Table 2 shows results of the GMM system in noisy conditions. The segmentation results are obtained from a full segmentation in each noise and SNR condition. We show frame selection results using the optimal EPC for each condition (shown in parentheses for the last system). The baseline EER in the clean condition was 9.09% (Table 1).

First, we observe in Table 2 a significant degradation under noisy conditions. We see that adding frame selection on top of segmentation is advantageous and results in additional EER reductions in all conditions. The optimal EPC for frame selection is dependent on both noise type and SNR. Finally, we observe that Wiener filtering results in a small EER reduction in babble noise down to 5dB SNR, and produces significant EER reductions across all SNR conditions for city traffic noise. Overall, the proposed techniques complement each other and yield incremental gains when combined.

5.2. MLLR-SVM System

In Table 3 we show the EER results of the MLLR-SVM system in both noise conditions. Frame selection was not used since it is incompatible with the phoneloop MLLR-SVM setup.

Comparing Tables 2 and 3, we observe that the MLLR-SVM system is much more affected by the babble noise than is the GMM system. Note that the MLLR-SVM system is based on a more detailed speech model (a phone recognition loop) that is potentially more affected by the speech-like components of babble noise. We also observe that Wiener filtering results in EER reductions in both noise types, but is more effective for city traffic noise.

5.3. Automatic EPC Parameter Selection

The next step in the GMM system was to automatically find an appropriate energy percent parameter based on an estimate of the SNR. With our current framework we can compute an accurate SNR estimate based on the speech/nonspeech segmenter output. We used the nonspeech regions to estimate the noise power. Then we used the traditional SNR formula with a clean speech power approximated as the noisy speech power minus the noise power. One very important advantage of this SNR estimator is that the noise is not assumed to be stationary.

Table 2: EER Results on Arabic Mixer Data with **GMM** System for Babble and City Traffic Noise at Multiple SNRs. Optimal Energy Percent Parameter in Parentheses

System	20dB	15dB	10dB	5dB	0dB
Babble noise					
Segmented	11.28	12.14	13.97	17.82	21.72
Segmented + frame selection	10.94	11.78	13.80	16.67	19.86
Wiener + seg + frame selection	10.30 (0.3)	11.45 (0.3)	13.13 (0.3)	16.15 (0.5)	20.03 (0.7)
City Traffic noise					
Segmented	13.47	14.97	18.01	21.55	24.75
Segmented + frame selection	12.79	13.97	15.83	19.02	23.06
Wiener + seg + frame select	11.61 (0.5)	12.61 (0.5)	13.97 (0.5)	17.00 (0.7)	21.33 (0.7)

Table 3: EER Results on Arabic Mixer Data with **MLLR-SVM** System for Babble and City Traffic Noise at Multiple SNRs

System	20dB	15dB	10dB	5dB	0dB
Babble noise					
Segmented	11.07	12.95	16.66	21.39	29.78
Wiener + seg	10.94	12.32	15.98	20.87	30.64
City Traffic noise					
Segmented	11.61	12.13	14.98	19.21	25.73
Wiener + seg	10.60	11.61	13.43	16.97	25.41

In the experiments we added noise to each Arabic database waveform with a randomly selected SNR between 25dB and 0dB from a uniform distribution. Next we estimated the SNR using the previously described estimator. Finally, we used the following EPC parameters given the SNR regions: >20dB: 0.3, 20dB-10dB: 0.5, <10dB: 0.7. These values were chosen based on the results in Tables 2 and 3. This assignment was used for both noisy conditions. In Table 4 we present the EERs using the true SNR and the estimated SNR to determine the EPC parameter.

Table 4: Automatic EPC Parameter Selection Results on Arabic Mixer Data with **GMM** System based on True and Estimated SNRs for Babble and City Traffic Noises

System	SNR	Babble	City Traffic
Wiener + seg		16.47	15.48
Wiener + seg + frame selection	True	14.31	14.78
Wiener + seg + frame selection	Estimated	14.48	14.65

From Table 4 one can conclude that the energy percent parameter selection performs about as well based on either the estimated or the true SNR.

5.4. Combiner

We were also interested in how the combiner described in Section 3.3 performs in noisy conditions. We tested the combiner using the same jackknifing procedure as described previously for clean speech. We tested two combiners. The first combiner was trained using scores in the matched noise and SNR condition. The second combiner was trained in clean conditions only. The same neural network model was used for testing on both noises and all SNRs. We also report in Table 5 for this last combiner the average of false acceptance (FA) and false rejection (FR) using the score threshold corresponding to the EER in clean conditions.

Table 5: EER Results on Arabic Database with **GMM, MLLR-SVM** Systems and **Combiners** for Babble and City Traffic Noises at Multiple SNRs. Parentheses: (FA+FR)/2 with Clean Threshold

System	20dB	15dB	10dB	5dB	0dB
Babble noise					
GMM	10.30	11.45	13.13	16.15	20.03
MLLR-SVM	10.94	12.32	15.98	20.87	30.64
Combiner Match Training	9.59	10.60	12.96	16.49	20.87
Combiner Train Clean Data	9.42 (8.7)	10.41 (9.9)	12.46 (11.3)	15.82 (17.4)	22.22 (30.8)
City Traffic noise					
GMM	11.61	12.61	13.97	17.00	21.33
MLLR-SVM	10.60	11.61	13.43	16.97	25.41
Combiner Match Training	9.27	9.76	11.11	14.31	20.32
Combiner Train Clean Data	9.42 (8.5)	9.62 (9.0)	11.11 (11.3)	15.02 (15.7)	19.86 (23.0)

We conclude from Table 5 that both combiners produce a gain over each system alone in city traffic noise and in babble noise up to 10dB SNR. Surprisingly, the clean combiner with the threshold obtained in clean conditions achieves similar results to the matched combiner for SNRs up to 5dB.

6. CONCLUSIONS

We explored two increasingly important challenges for speaker recognition applications: noise robustness and porting to a language of interest (Arabic). We described a noise robust speaker identification system that includes multiple components and multiple models. We found complementary gains from the multiple noise robust components, especially in combination. In city traffic noise we obtained a gain over all SNRs by using score combination. Interestingly, we found that the combiner trained in clean conditions performed similarly to one trained in matched conditions, a useful finding since matched data are often not available or practical. Furthermore, our systems were not specifically tuned for Arabic. For example, the MLLR-SVM system used English acoustic models. Thus, future work in which such systems are tailored to the language might yield additional performance gains.

7. ACKNOWLEDGMENTS

This work was funded by awards NMA401-02-9-2001 and NSF IIS-0544682. The views herein are those of the authors and do not reflect the views of the funding agencies.

8. REFERENCES

[1] J. Ming, T. Hazen, and J. Glass, "A Comparative Study of Methods for Handheld Speaker Verification in Realistic Noisy Conditions," *Proc. of IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.

[2] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Speaker Identification under Noisy Environments by Using Harmonic Structure Extraction and Reliable Frame Weighting," *Proc. of International Conference on Spoken Language Processing (Interspeech-2006)*, pp. 1459-1462, USA, 2006.

[3] S. Ouamour-Sayoud, H. Sayoud, and M. Boudraa, "Application of the MLVQ1 in Speaker Identification," *ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP'03)*, France, 2003.

[4] http://www.nist.gov/speech/tests/spk/index.htm

[5] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Mixture Models," *Digital Signal Processing*, vol. 10, pp.181-202, 2000.

[6] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman "MLLR Transforms as Features in Speaker Recognition," *Proc. Eurospeech, Lisbon*, pp. 2425-2428, 2005.

[7] MIT Lincoln Laboratory, LNKNet, http://www.ll.mit.edu/IST/lnknet/

[8] L. Ferrer, E. Shriberg, S. S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman, and H. Bratt, "The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System," *Proc. IEEE ICASSP*, vol. 1, pp. 101-104, Toulouse, 2006.

[9] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," *Proc. ICSLP*, vol. 1, pp. 4-7, Denver, Sep. 2002.

[10] NOISEX database samples available at http://spib.rice.edu/spib/select_noise.html