ADVANCES IN THE CMU/INTERACT ARABIC GALE TRANSCRIPTION SYSTEM

Mohamed Noamany, Thomas Schaaf[*]*, Tanja Schultz*

InterACT, Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213
{mfn,tschaaf,tanja@cs.cmu.edu}

ABSTRACT

This paper describes the CMU/InterACT effort in developing an Arabic Automatic Speech Recognition (ASR) system for broadcast news and conversations within the GALE 2006 evaluation. Through the span of 9 month in preparation for this evaluation we improved our system by 40% relative compared to our legacy system. These improvements have been achieved by various steps, such as developing a vowelized system, combining this system with a non-vowelized one, harvesting transcripts of TV shows from the web for slightly supervised training of acoustic models, as well as language model adaptation, and finally fine-tuning the overall ASR system.

Index Terms— Speech recognition, Vowelization, GALE, Arabic, Slightly supervised training, web data.

1. INTRODUCTION

The goal of the GALE (Global Autonomous Language Exploitation) program is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages and make them available in English. In a long run this requires to combine techniques from text summarization, information retrieval, machine translation, and automatic speech recognition. NIST will perform regular evaluations and the first evaluation took place recently. This paper describes improvements in the CMU Modern Standard Arabic (MSA) system through the span of 9 months in preparation for this evaluation.

One of the language characteristics and challenges of Arabic is that some vowels are omitted in the written form. These vowels carry grammatical case information and may change the meaning of a word. Modeling the vowels in the pronunciation dictionary was found to give improvements over un-vowelized pronunciations [4]. In this paper we achieved another significant improvement by combining a vowelized with a non-vowelized system. Furthermore, we got gains by collecting and utilizing web transcripts from TV show, which include broadcast conversations.

2. SYSTEM DESCRIPTION

Our MSA speech recognition system is based on the Janus Recognition Toolkit JRTk [9] and the IBIS decoder [10].

Before decoding the audio, an automatic segmentation step and a speaker clustering step is performed. The segmentation step aims at excluding those segments that contain no speech, such as music or background noise. The remaining segments are clustered into speaker clusters such that all adaptation and normalization steps can be processed on clusters as batches.

From the incoming 16 kHz audio signal we extract for each segment power spectral features using a FFT with a 10ms frame-shift and a 16ms Hamming window. From these we compute 13 Mel-Frequency Cepstral Coefficients (MFCC) per frame and perform a cepstral mean as well as variance normalization on a cluster basis. To incorporate dynamic features we concatenate 15 adjacent MFCC frames (±7) and project these 195 dimensional features into a 42 dimensional space using a transform found by linear discriminate analysis (LDA). We use the context-dependent codebooks as classes for finding the LDA transform [2]. On top of the LDA we apply a single maximum likelihood trained Semi-Tied-Covariance (STC) matrix.

The general decoding setup employs a first pass in which a speaker independent acoustic model without vocal tract length normalization (VTLN) and no adaptation is used. The hypotheses of a cluster from the first pass are then used to estimate the VTLN warping factors to warp the power spectrum using the maximum likelihood approach described in [8]. After the VTLN factors are found, the same hypotheses are considered to estimate a feature space adaptation (FSA) using a constrained MLLR (CMLLR) transform. Then a model space adaptation is performed using maximum likelihood linear regression with multiple regression classes. The regression classes are found through clustering of the Gaussians in the acoustic model. The second pass decoding uses a speaker adaptive trained

[*] Now with Toshiba Research Europe Ltd, Cambridge, United Kingdom

Proceedings of NAACL HLT 2007, Companion Volume, pages 129–132,
Rochester, NY, April 2007.

acoustic model, in which the adaptation was performed using a single CMLLR transform per speaker.
For the non-vowelized system, we applied a grapheme-to-phoneme approach to automatically generate the pronunciation dictionary. For the vowelized system we used the same phoneme set as in the non-vowelized system but extended it with the 3 short vowels, which do not appear in the writing system. Both systems are 2-pass system as described above and employ Cepstral Mean Normalization (CMN), MLLR, Semi-tied covariance (STC), and Feature space adaptation (FSA).

For the development of context dependent acoustic models we applied an entropy-based polyphone decision tree clustering process using context questions of maximum width ±2, resulting in shared quin-phones. In addition we included word-boundary tags into the pronunciation dictionary, which can be asked for in the decision tree can ask for word-boundary tags. The non-vowelized system uses 4000 phonetically-tied quin-phones with a total of 305,000 Gaussians. The non-vowelized system has 5000 codebooks with a total of 308,000 Gaussians.

In total we used 190 hours for acoustic training. These consist of 40 hours Broadcast news (BN) from manually transcribed FBIS data, 50 hours BN LDC-TDT4 selected from 85 hours using a slightly supervised approach as described in [3], and 30 hours Broadcast conversation (BC) recorded from Al-jazeera TV, and 70 hours (40hrs BN, 30hrs BC) from LDC-GALE data. For quality reasons we removed some of the most recent GALE data from acoustic model training.

4. LANGUAGE MODELING

The Arabic Giga word corpus distributed by LDC is currently the major Arabic text resource for language modeling. Since this corpus only covers broadcast news, we spidered the web to cover broadcast conversational data. We found transcripts for Arabic talk shows on the Al-jazeera web site www.al-jazeera.net and collected all data available from 1998 to 2005. We excluded all material from 2006 to comply the evaluation rules which prohibit the use of any data starting February 2006. In addition to the mentioned data we collected BN data from the following source: Al-Akhbar (Egyptian daily newspaper 08/2000 to 12/2005) and Akhbar Elyom (Egyptian weekly newspaper 08/2000 to 12/2005). Furthermore, we used unsupervised training transcripts from 750 hours BN created and shared by IBM.

For language modeling building we used the SRILM tool kit from SRI [5]. Since we have 2 kinds of data, Broadcast News and Conversation, we built various individual 4-grams language models. 11 models were then interpolated to create one language model. The interpolation weights were selected based on a held out data set from BN and BC sources. We found that the data from Al-jazeera (both BN & BC) has the highest weight comparing to other sources. The resulting final language model uses a total number of n-grams is 126M and a vocabulary of 219k words. The perplexity of the language model is 212 on a test set containing BC and BN data.

5. TV WEB TRANSCRIPTS

Most of our acoustic and language model training data comes from broadcast news. However, since GALE targets broadcast news as well as conversations we looked for an effective method to increase the training data for Arabic BC. We made use of the fact that some Arabic TV stations place transcripts for their program on the web. These transcripts lack time stamp but include acceptable quality of the transcription. However, one challenge is that the transcriptions are not complete in that they do not include transcripts of commercials or any news break that may interrupt the show. In total we recorded 50 hours of Broadcast conversation shows from Al-jazeera and used them in our acoustic model and language model training by performing the following procedures:

- We manually selected shows from Al-jazeera TV
- We used a scheduler to automatically start the recording of the selected shows.
- We spidered the web to collect corresponding show transcripts from their web site www.aljazeera.net.
- We automatically processed the transcripts to convert the html files to text, convert numbers to words and remove any non-Arabic words in the shows.
- We added these shows to our LM data with high weight, built a biased LM, and used this LM to decode the recorded shows.
- We aligned the reference (transcripts without time stamps) with the decoder output that may contain speech recognition errors.
- We selected only the portions that are correct; we did not select any portion with number of words less than 3 correct consecutive words.
- Based on the above criteria we finally selected 30 hours out of the total 40 hours recordings.
- We clustered utterances based on BIC criteria approach described in [7].

As a result, we managed to project the time stamp in the original transcript such that it can be used for training. Using these 30 hours of data resulted in a 7% relative improvement on RT04. Since RT04 is broadcast news, we expect even higher gains on broadcast conversational data. It is worth mentioning that we

applied the same slightly supervised approach to the TDT4 data which is a low quality quick transcription. We selected 50 out of 80 hours and achieved an improvement of 12% relative. The gain was higher since at the time of these experiments we had only 40 hours of training from FBIS data, therefore more than doubled the amount of training data by adding TDT4.

6. NON-VOWELIZED SYSTEM

Arabic spelling is mostly phonemic; there is a close letter-to-sound correspondence. We used a grapheme-to-phoneme approach similar to [1]. Our phoneme set contains 37 phonemes plus three special phonemes for silence, non-speech events, and non-verbal effects, such as hesitation.

We preprocessed the text by mapping the 3 shapes of the grapheme for glottal stops to one shape at the beginning of the word since these are frequently miss-transcribed. This preprocessing step leads to 20% reduction in perplexity of our language model and 0.9% improvements in the final WER performance on RT04. Preprocessing of this kind appears to be appropriate since the target of the project is not transcription but speech translation and the translation community applies the same pre-processing. We used a vocabulary of 220K words selected by including all words appearing in the acoustic transcripts and the most frequent words occurring in the LM. The OOV rate is 1.7% on RT04. Table 1 shows the performance of our Speaker-Independent (SI) and Speaker-Adaptive (SA) non-vowelized system on the RT04 set.

Table 1: Non-vowelized System Results

System		WER on RT04 (%)
Non-Vowelized	(SI)	25.3
Non-Vowelized	(SA)	20.8

7. VOWELIZED SYSTEM

Written MSA lacks vowels, thus native speakers add them during reading. Vowels are written only in children books or traditional religious books. To restore vowels for a 129K vocabulary [4], we performed the following steps:

- Buckwalter morphological analyzer (BMA) (found 106K out of 129K entries).
- If a word is not vowelized by the analyzer, we check for its vowelization in the LDC Arabic Tree-Bank (additional 5k entries found).
- If the word did not appear in any of those, we used the written non-vowelized word form.

In total 11k entries could not be resolved by either the BMA or the Treebank.

This vowelization step resulted in 559,035 pronunciations for the 129k words in our vocabulary, i.e. we have on average 5 pronunciations per word. To reduce the number of pronunciation variants we performed a forced alignment and excluded pronunciations which did not occur in the training corpus. This results in 407,754 pronunciations, which is a relative reduction of about 27%. For system training we used the same vocabulary and applied the same training procedure as in the non-vowelized system for acoustic model training.

As Table 2 shows, we achieved a very good gain of 1.3% absolute on the SI pass and 1.5% on the SA pass, both benchmarked on RT04 (compare Table 1). We envision to seeing even higher improvements after estimating and applying probability priors to multiple pronunciation and after vowelizing the remainder 11k words that had not been covered by BMA or the Tree-Bank.

Table2: Vowelized System Results

System		WER on RT04 (%)
Vowelized	(SI)	24.0
Vowelized	(SA)	19.3

8. COMBINING VOWELIZED & NON-VOWELIZED SYSTEM

After seeing significant improvements by vowelization, we investigated the performance gain through cross-adapting the vowelized system with the non-vowelized system. The vowelized system cross adapted with the SA non-vowelized gave us 1.3 over the vowelized system adapted on the SI vowelized system. We used a 3-pass decoding strategy, in which the first pass uses the speaker independent (SI) vowelized system, the second pass uses the speaker adaptive (SA) non-vowelized system, and the third, final pass, uses the speaker adaptive vowelized system. Some challenges for the cross-adaptations had to be overcomed, for instance to cross adapt the non-vowelized system on the vowelized system, we had to remove the vowels to have a non-vowelized transcript. Since the phoneme set of the non-vowelized system is a subset of the phoneme set of the vowelized system, we could simply exclude the vowel phonemes from the vowelized system. Furthermore, the search vocabulary is the same and so is the language model.

The main changes are the pronunciation dictionary and the decision tree. We tried different combination schemes, e.g. by starting with the non-vowelized system, then the vowelized, and then the non-vowelized but found that none outperforms the combination reported here in terms of WER. In addition starting with the non-vowelized SI pass is much faster than the vowelized SI system (4.5RT compared to 9RT).

Table 3: Non-vowelized & vowelized System Combination

System	WER on RT04 (%)
Vowelized (SI)	24.0
Non-Vowelized (SA)	19.9
Vowelized (SA)	18.3

9. ACOUSTIC MODEL PARAMETER TUNING

We started our legacy system with 40 hours and until it reached 90 hours we were using the same number of codebooks (3000) and same number of Gaussians (64) per codebook. With the increase of training data from 90 hours to 190 hours we investigated the effect of increasing the number of codebooks and Gaussians. Also, we were using merge and split training (MAS) and STC only for the adapted pass; we furthermore investigated the effect of using it for the SI pass. We found that using MAS & STC on the SI pass gave us a gain of 5% relative on the SI pass. In addition we found that the ideal number of codebooks is 5000 for the non-vowelized system resulting in a gain of 5.3% relative on the SI pass. We expect to see further gains on the SA pass. Table 4 summarizes the system performance using different parameter sizes and training schemes.

Table 4: System Performance vs.Model Size

#codebooks	MAS	#Gausians	Voc	System	WER(%)
3K	-	64K	129	Non-vow(NV)	29.6
3K	***Mas***	64K	129	NV	28.3
5K	Mas	64K	129	NV	27.9
5K	Mas	***100K***	129	NV	27.6
5K	Mas	100K	200	nv+***tv TRANS***	26.3
3K	Mas	100K	200	***vow***+ tvTRANS	24.0

10. SYSTEM EVOLUTION

Table 5 shows the gains we achieved at major milestone stages while building the system. The key improvements are due to adding data collected from the web, Vowelization, and combining the vowelized and non-vowelized systems. Tuning the acoustic models parameters gave us a good gain and finally the interpolation of different language model for different sources gave additional improvements. The real-time behavior of the system improved from 20RT to 10 RT while loosing only 0.2% which is in acceptable trade-off. Recently, we gained 3.5% relative applying discriminative training (MMIE).

11. CONCLUSION

We presented the CMU 2006 GALE ASR Arabic system. It can be seen that we achieved 40% improvements over our legacy system.

Table 5: System Progress WER (%)

LEGACY SYSTEM	32.7
STC+VTLN	30.1
SPEED FROM 20RT TO 10RT	30.3
FROM 3 TO 4GM+BETTER SEGMENTATION	28.4
TDT4 TRANSCIPTS SELECTION REFINEMENT	26.3
CLUSTERING REFINEMENT & RETRAINING	25.5
MORE LM DATA +INTERPOLATING 11 LMS	24.2
ADDITION Q3 OF LDC DATA	23.6
ACOUSTIC MODEL PARAMETER TUNING	20.7
MMIE	20.0
COMBINED SYSTEMS (VOW+NON-VOW)	18.3

We combined a vowelized and a non-vowelized system and achieved 4.0% relative over the vowelized system. Also, we managed to use TV web transcript as a method to cover the shortage of training data specially the broadcast conversation. Currently, we are exploring more on the vowelized system by adding weights to different multiple pronunciations and adding vowelization to words not covered by the morphological analyzer or the tree-bank.

12. ACKNOWLEGMENTS

We also would like to thank Qin Jin for applying her automatic clustering techniques to the web data.

13. REFERENCES

[1] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, "Audio Indexing of Arabic Broadcast News", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2002.

[2] H. Yu, Y-C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz "The ISL RT04 Mandarin Broadcast News Evaluation System", EARS Rich Transcription workshop, Palisades, NY, 2004.

[3] L. Nguyen et al., "Light supervision in acoustic model training," ICASSP , Montreal, QC, Canada, May 2004.

[4] M. Afify et al.,"Recent progress in Arabic broadcast news transcription at BBN", In INTERSPEECH-2005.

[5] A. Stolcke SRILM- An Extensible Language Modeling ToolKit ICSLP. 2002, Denver, Colorado.

[6] T.Buckwalter, "Issues in Arabic Orthography and morphology Analysis", COLING 2004, Geneva, 2004.

[7] Q. Jin, T. Schultz, "Speaker segmentation and clustering in meetings", ICSLP, 2004.

[8] P. Zhan, M. Westphal, "Speaker Normalization Based On Frequency Warping", ICASSP 1997, Munich, Germany.

[9] M. Finke, et al., "*The Karlsruhe Verbmobil Speech Recognition Engine*," International Conference on Acoustics, Speech, and Signal Processing, ICASSP,1997.

[10] H. Soltau, et al., "A One Pass-Decoder Based On Polymorphic Linguistic Context", ASRU 2001, Trento, Italy, 2001