Recent Progress in Arabic Broadcast News Transcription at BBN

Mohamed Afify, Long Nguyen, Bing Xiang, Sherif Abdou, and John Makhoul

BBN Technologies
10 Moulton Street, Cambridge, MA, 02138, USA
{mafify,ln,bxiang,makhoul}@bbn.com

ABSTRACT

The first part of this paper describes the BBN system that participated in the 2004 broadcast news (BN) evaluation for Arabic. The complete system description is given together with experimental results on the 2004 development, and evaluation sets. Previous Arabic speech recognition at BBN used grapheme models due to the lack of short vowel information in the acoustic transcriptions. In the second part of this paper we show how to build a phonetic system. It is demonstrated that switching to phonetic models is capable of reducing the word error rate by up to 14% relative, for different test sets, compared to the traditional grapheme based approach.

1. Introduction

This paper first describes the BBN broadcast news (BN) transcription system that was submitted to the 2004 EARS evaluation. A detailed system description is given, and experimental results for the evaluation system are also provided. Previous Arabic speech recognition systems developed at BBN, including the 2004 evaluation system, used grapheme acoustic models. The main reason is the absence of short vowels from most available Arabic texts. This makes phonetic transcription, and hence building phonetic models, a difficult and time consuming task. Thus, the grapheme approach provided a rapid way to construct Arabic speech recognition systems with reasonable performance.

In recent work [4], and also in the 2004 EARS evaluation, it was demonstrated that Arabic phonetic models are capable of providing superior results to the traditional grapheme based systems. Thus, the second part of this paper discusses our efforts to construct a phonetic system for Arabic BN. The focus is on incorporating short vowel information in the acoustic model to build a phonetic system. We, thus, turn our attention to the rapid construction of short vowel information that is usually missing from acoustic transcripts, and show how to quickly bootstrap phonetic acoustic models. Performance of the resulting system is compared, on different test sets, to traditional grapheme based models, and show improvements up to 14% in word error rate.

The paper is organized as follows. First we give general architectures that are used in both systems in Section 2. Section 3 gives an account on acoustic, and language model training data used in both systems. This is followed by Sections 4, and 5 which describe both the evaluation system, and the phonetic system, and give experimental results on different development and evaluation sets. Finally, the paper is concluded in Section 6.

2. System Architecture

The BBN 2004 evaluation system for Arabic BN transcription consists of a combination of two systems using ROVER [1]. The two systems are configured to run under $10\times$RT, and are referred to as **B1**, and **B2** respectively. Both systems will be described below, and we refer the reader to [6], and references therein, for more details. Also the Arabic phonetic system that we built is similar in structure to **B1**, and hence this section will serve in introducing this system as well.

2.1. Architecture of the First System (B1)

The system comprises audio segmentation, feature extraction, and decoding.

The input audio is first segmented [6] based on gender. For each gender, speaker change detection is performed using the Bayesian Information Criterion (BIC). The speaker turns are then chopped into short segments, averaging 7 seconds, according to silence locations. The resulting segments are clustered using an online algorithm that employs a penalized likelihood measure. The obtained clusters are used for adaptation and decoding.

After the segmentation and clustering stage, features are extracted from overlapping frames of speech of length 25 ms at a rate of 100 frames/sec. For each frame, 14 perceptual linear prediction (PLP) cepstral coefficients, and energy are calculated. The static features are augmented with their first, second, and third order derivatives, which leads to an initial 60-dimensional parameter space. The dimension is then reduced to 46 using heteroscedastic linear discriminant analysis (HLDA).

Following feature extraction comes decoding. The decoding consists of two stages: the unadapted, and adapted decoding stages. The first stage uses speaker independent (SI) models, and provides supervision for model adaptation in the second stage that employs speaker adaptively trained (SAT) models. The decoding strategy is common to the two stages, and will be outlined below.

Decoding employs a multi-pass search startegy, where each pass is used to constrain the search space of the following pass. In the current system, a forward pass and a backward pass are run followed by N-best rescoring. These will be described briefly below.

- The forward pass uses simple acoustic models, Phonetically Tied Mixture (PTM) models, and a bigram language model. The outputs are the most likely words at each frame together with their scores.
- The backward pass then uses the output of the forward pass to guide a Viterbi beam search with more complex acoustic and language models. A state clustered (using decision trees) within-word quinphone acoustic model (SCTM-NX), and an approximate trigram language model are used in this step. During the backward pass an N-best list is generated.
- In the current system, an N-bset list (N=300) is output by

the backward pass. This list is rescored using the SCTM-NX model[1], and a 3-gram language model. The top scoring hypothesis represents the recognition output.

Adaptation in the second pass starts by estimating a speaker specific HLDA transform [3] for each speaker, followed by applying a constrained maximum likelihood linear regression (CMLLR) transform [2]. The final adaptation step, after the above two feature space transforms, amounts to estimating two MLLR transforms of the model parameters, based on a tree clustering of the model distributions. MLLR is applied to both the PTM and the SCTM-NX models.

2.2. Architecture of the Second System (B2)

This system also comprises audio segmentation, feature extraction, and decoding. It starts with the audio segmentation and clustering (see for example Section 2.1 above) followed by feature extraction. 14 PLP cepstral coefficients and energy are calculated for each frame, and appended by their first, and second order derivatives. This results in a 45-dimensional feature space.

The decoding, as for system **B1**, consists of two stages. The un-adapted stage ouputs a transcription which is used as supervision to adapt the models. Finally the adapted models are used for the second adapted stage. Both stages employ gender dependent (GD) acoustic models, and hence no SAT model is trained for the adaptation pass. The GD models are trained using maximum a posteriori (MAP) estimation, starting from a speaker independent model.

A multi-pass search startegy, similar to system **B1**, is employed. This consists of a forward pass which uses a PTM acoustic model and a bigram language model. The forward pass is followed by a backward pass which uses an SCTM-NX acoustic model, and an approximate trigram language model, and generates an N-best list of size 300. The final step rescores the N-best using the SCTM-NX model and a trigram language model. The adaptation in this case does not use any feature space transformations, and only two MLLR transforms are used for model adaptation. Table 1 shows the most important differences between systems **B1**, and **B2**.

	B1	**B2**
Feature size	46	45
HLDA	yes	no
SAT model	yes	no
GD model	no	yes
CMLLR	yes	no
Passes	2	2

Table 1: Most important differences between the two components used in the Arabic BN system for 2004 evaluation.

3. Training and Testing Data

This section describes the acoustic model, and language model training data, the pronunciation lexicons, and the test sets used in the evaluation, and phonetic systems.

[1]Typically cross-word models are used in rescoring but they were found to lead to worse performance for Arabic.

1. **Acoustic Model Training Data:** The 2004 system acoustic training data consists of about 100 hours. These include 72 hours from the TDT data for Arabic, that are processed by light supervision [5], and the rest from FBIS. We refer to this as the full training (**FT**) set. We also use a subset of the data of size 80 hours in training the phonetic system. This will be referred to as the reduced training (**RT**) set. The transcriptions of **RT** are text normalized as will be discussed in Section 5.
2. **Language Model Training Data:** We selected about 300M words for building the language model. These consist of several news paper text, and web data in Arabic. As will be discussed in Section 5 this data was also text normalized. We refer to both the original, and normalized data sets as **UD**, and **ND**, respectively. Both sets have the same size of 300M words.
3. **Pronunciation Lexicons:** An initial pronunciation lexicon of size 64K words was chosen using the **UD** set. After text normalization, this lexicon was reduced in size to about 62K words. Due to discarding some words in the vowelization process, as will be discussed in Section 5, the lexicon size was further reduced to 60K words. We refer to the 64K, and 60K lexicons as the full lexicon **FL**, and the reduced lexicon **RL** respectively.
4. **Test Sets:** Four test sets are used in evaluating the systems. Namely, the 2003, and 2004 development, and evaluation sets. We refer to these sets as Dev03, Dev04, Eval03, and Eval04 respectively. For the phonetic system, all the references are text normalized in the same way as the acoustic model transcriptions, and language model data.

The characteristics of the acoustic, and language model data sets, and the pronunciation lexicons are summarized in Table 2.

	Type	Size	Normalization
FT	AM data	100 hours	No
RT	AM data	80 hours	Yes
UD	LM data	300M words	No
ND	LM data	300M words	Yes
FL	Lexicon	64K words	No
RL	Lexicon	60K words	Yes

Table 2: The characteristics of the full acoustic training set **FT**, the reduced acoustic training set **RT**, the un-normalized language model training data **UD**, the normalized language model training data **ND**, the full pronunciation lexicon **FL**, and the reduced pronunciation lexicon **RL**. AM, and LM stand for acoustic, and langugae model respectively.

4. The Evaluation System

This section describes the acoustic and language models, and gives experimental results for the evaluation system. The evaluation system consists of the combination of two component systems called **B1**, and **B2**.

The two systems use the same grapheme set, and the same pronunciation lexicon **FL**. Each system uses hidden Markov models (HMMs) to represent an inventory of 39 graphemes, where each

System	Model	# Gaussians
B1	PTM SI	19K
B1	SCTM-NX SI	229K
B1	PTM SAT	19K
B1	SCTM-NX SAT	147K
B2	PTM	19K
B2	SCTM-NX	365K

Table 3: Model size, for systems **B1** and **B2**, in number of Gaussians for the PTM, and SCTM-NX models used in the Arabic evaluation system.

HMM has a left-to-right topology and consists of 5 states. These grapheme models are used as building blocks for words in the 64K pronunciation lexicon. There are a total of 195 (39*5) states. A decision tree is built for each state, and each tree leaf is represented by a Gaussian mixture model.

System **B1** uses four acoustic models in total. Namely, the PTM, and the SCTM-NX for both SI, and SAT. For system **B2** also four acoustic models are needed for both decoding stages. These are the PTM, and the SCTM-NX for both genders. Only the SCTM-NX SAT model for system **B1** is trained using MMI while all other models are trained using ML. The models are trained using the **FT** training set of size 100 hours. Table 3 shows the number of Gaussians of each model for both systems. It should be noted that only two entries are shown for system **B2** because gender dependent models are obtained using MAP adaptation of the SI system, and hence have the same number of Gaussians for each gender.

A trigram language model is built using the **UD** training set of size 300M words. The system vocabulary is 64K corresponding to the pronunciation lexicon **FL**. The trigram model has about 67M trigrams, and 23M bigrams. The results for both Dev04, and Eval04 for both component systems, and the combined output are shown in Table 4. It should be noted that for the results shown both "tanween", and initial "hamza" normalization are applied as implemented by NIST.

Set	**B1**	**B2**	ROVER
Dev04	19.0	19.6	17.4
Eval04	22.1	24.5	21.9

Table 4: Word error rate on the Arabic 2004 development and evaluation sets for systems **B1** and **B2**, and the system combination result is also shown in the column labeled ROVER.

5. The Phonetic System

Short vowels in Arabic are diacritics placed above or below the letters, and are usually missing from most available Arabic texts. These short vowels are necessary to perform phonetic transcription in Arabic, and hence to build phonetic acoustic models. In this section we show how we can supply the missing short vowel information for acoustic transcripts, and provide details for building an Arabic phonetic system.

5.1. Bootstrapping Short Vowels for Transcriptions

The missing short vowels are added to the acoustic transcriptions, and the pronunciation lexiocn using two resources available from the linguistic data consortium (LDC): the Buckwalter morphological analyzer, and the Arabic treebank corpus.

The Buckwalter morphological analyzer outputs possible vowelizations of a word that is in its dictionary. It usually only misses foreign, or mis-spelled words. In our initial development we used version 1.0 of the morphological analyzer, then we switched to version 2.0 upon its release. The Arabic Treebank corpus consists of vowelized news articles. From these articles a dictionary that contains each word and its possible vowelizations can be formed. Using these two resources our vowelization method is very simple:

- Pass a word to the Buckwalter morphological analyzer, and assign to it all the ouput vowelizations.
- If the word is missed by the analyzer, look it up in the Treebank dictionary, and ouput all possible vowelizations.
- If both steps fail either discard the word, or manually vowelize it.

The 100 hours **FT** training set has 50K unique words. Our initial goal was to vowelize these words in addition to the 64K words in the **FL** pronunciation lexicon. Our vowelization procedure misses some words. Thus, in order to limit the manual effort we discarded any training sentence that has any unvowelized words, and also all unvowelized words in the pronunciation lexicon. In our initial system development, we noticed that the phonetic system is very sensitive to text normalization compared to the grapheme system. For this reason we applied a simple normalization procedure to acoustic, and language model data, and reference transcripts of the test data. The normalization simply maps all forms of the "hamza" at the beginning of the word, and after popular prefixes to the letter "Alef", and also corrects some very frequent confusions of the letters "Y", and "y" at the end of the word. This way we retained about 80 hours of net acoustic training with normalized transcriptions, a pronunciation lexicon of about 60K words, and 300M words of normalized LM data. These were referred to in Section 3 as **RT**, **RL**, and **ND** respectively. These will be used in building the phonetic system, and also a grapheme system for comparison purpose[2].

Once both the training data, and the dictionary are vowelized, it is straightforward to perform phonetic transcription in Arabic. This is basically a one-to-one mapping with very few exception rules. The phonetic set consists of 35 phonemes (28 consonants, 6 vowels, and "taa marbUTa"), in addition to silence, garbage, and hesitation symbols. This is to be compared to 39 sysmbols, which include the same non-speech symbols, for the grapheme system. After deciding on the phonetic set, and performing phonetic transcription, it is straightforward to build a phonetic system in the same way as the grapheme system. This will be discussed below.

5.2. Phonetic System Development and Results

The phonetic system is similar in structure to system **B1** in Section 2.1. The only differences are that we build 5-state HMMs

[2]Note that the evaluation system is not directly comparable to the phonetic system due to the difference in amount of training data, size of pronunciation lexicon, and text normalization.

Model	Grapheme	Phonetic
PTM SI	18K	18K
SCTM-NX SI	183K	183K
PTM SAT	18K	18K
SCTM-NX SAT	116K	116K

Table 5: Model size, for the grapheme and phonetic systems, in number of Gaussians for the PTM, and SCTM-NX acoustic models.

to represent 38 phonetic sysmbols instead of the 39 graphemes. In addition, no MMI estimation is used, and all models are built using ML estimation. The model is trained using the 80 hours set **RT**. For comparison purpose, we also build a grapheme system using the same 80 hours of acoustic training data. The size, in number of Gaussians, of the PTM, and SCTM-NX for both SI, and SAT are shown in Table 5 for both systems.

As stated above, the use of short vowels is limited to the acoustic model. The normalized LM training data set **ND**, and the 60K vocabulary in **RL** are used in building the language model. When using version 1.0, and 2.0 of the Buckwalter morphological analyzer there are about 113K, and 300K different pronunciations corresponding to these 60K words, respectively. The resulting trigram language model has 64M trigrams, and 22M bigrams, which is comparable to that of the evaluation system.

For the purpose of development we use the Dev03 test set. We also present results for only one decoding pass, i.e. unadapted decoding. These initial results are shown in Table 6.

System	Morphological Analyzer	WER
Grapheme	NA	18.1
Phonetic	Version 1.0	16.9
Phonetic	Version 2.0	15.8

Table 6: Word error rate for unadapted decoding on the Arabic 2003 development set for the grapheme, and phonetic systems. Both versions 1.0 and 2.0 of the Buckwalter morphological analyzer are tested for vowelization.

As can be observed from the table, both phonetic systems outperform the grapheme system, and the phonetic system using version 2.0 of the Buckwalter morphological analyzer also outperforms that using version 1.0 of the analyzer. The main difference between the two versions is that version 2.0 of the analyzer generates all possible ending vowelizations, including "tanween", of a word. The best phonetic system gives about 13% reduction in WER compared to the grapheme system. Due to these encouraging results, we performed two pass decoding (including adaptation) on different test sets. These results will be presented below.

The adapted decoding results on the test sets Dev03, Dev04, Eval03, and Eval04 are given in Table 7. The phonetic system reported here uses version 2.0 of the Buckwalter morphological analyzer due to its clearly better results.

As can be observed from the table, there is a reduction from 10%-14% on all the test sets by switching to the phonetic system. This

Test Set	Grapheme	Phonetic
Dev03	15.4	14.2
Dev04	18.9	16.8
Eval03	18.9	16.3
Eval04	23.4	21.1

Table 7: Word error rate on the Arabic 2003 , and 2004 development and evaluation sets for the grapheme, and phonetic systems.

can be explained by the sharper acoustic models resulting from modeling phonemes rather than graphemes. This is further supported by the fact that both systems have almost the same number of Gaussians as shown in Table 5.

6. Conclusion

In the first part of this paper, we presented the BBN 2004 evaluation system for Arabic BN. We outlined the general structure, followed by experimental results on the 2004 development and evaluation sets. In the second part of the paper, we considered switching to a phonetic system instead of using grapheme acoustic models. In particular, we showed how to quickly provide short vowel information for the acoustic transcriptions, and hence build a phonetic system. The phonetic system shows a consistent improvement of 10%-14%, for different test sets, over a similar grapheme system. Future research will focus on methods to automatically vowelize the missing words, and hence to make full use of all the available acoustic data, and also to use a larger pronunciation lexicon. We will also explore introducing vowelization in the language model in addition to the acoustic model.

References

1. J. G. Fiscus. "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," IEEE ASRU Workshop, 1997.
2. M. J. F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition," Tech. Report CUED/F-INFENG/TR291, Cambridge University Engineering Dept., 1997.
3. S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections", Automatic Speech Recognition and Understanding, 2003, St. Thomas, Dec. 2003.
4. A. Messaoudi, L. Lamel, and J.L. Gauvain, "Transcription of Arabic broadcast news," in Proc. ICSLP'04, Jeju Island, Korea, October 2004.
5. L. Nguyen, B. Xiang, "Light Supervision in Acoustic Model Training," in Proceedings of ICASSP'04, Montreal, May 2004.
6. L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, and J. Makhoul, "Progress in Transcription of Broadcast News Using Byblos," Speech Communication, 38:213-230, 2002.