

RECENT IMPROVEMENTS TO THE CAMBRIDGE ARABIC SPEECH-TO-TEXT SYSTEMS

M. Tomalin, F. Diehl, M.J.F. Gales, J. Park & P.C. Woodland

Cambridge University Engineering Department, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {mt126, fd257, mjfg, jhp33, pcw}@eng.cam.ac.uk

ABSTRACT

This paper describes recent improvements to the Cambridge Arabic Large Vocabulary Continuous Speech Recognition (LVSCR) Speech-to-Text (STT) system. It is shown that Multi-Layer Perceptron (MLP) features trained on phonetic targets can improve the performance of both phonemic and graphemic systems. Also, a morphological decomposition scheme is extended from the graphemic domain to the phonetic domain, and particular attention is given to the task of dictionary generation. Finally, the use of Boosted Maximum Mutual Information (BMMI) training is explored both for individual systems and in the context of system combination. The full system results show that the combined use of the above techniques reduces the Word Error Rate (WER) of the best individual system by up to 12% relative, and that the incorporation of morphological decomposition and BMMI within the four individual branches of the combined system reduces the WER by up to 9% relative.

Index Terms— Arabic, Speech-to-Text, MLP features, Morphological Decomposition, Boosted MMI

1. INTRODUCTION

It is well-known that Arabic poses non-trivial problems for state-of-the-art STT systems: the language is morphologically complex; the Arabic dialects differ considerably, and Modern Standard Arabic (MSA) is conventionally written without vowels. To overcome this final difficulty, graphemic and phonetic Arabic STT systems are often constructed: the former are trained using unvowelised data, while the latter use vowelised data, and the output from the different types of systems is subsequently combined [1, 2].

This paper describes the recent performance gains that have been obtained for the Cambridge Arabic STT system, and the discussion is structured as follows. Section 2 describes the baseline system and explores the use of MLP features in Arabic STT systems. Particular attention is given to the use of phonetic MLP training targets for both phonetic and graphemic systems. Section 3 focuses on the impact of morphological decomposition schemes for both graphemic and phonetic systems. In section 4, the use of Boosted Maximum Mutual Information (BMMI) training is considered in the context of system combination with Minimum Phone Error (MPE) trained systems. When these three techniques are implemented, the WER for the best single system is reduced by up to 1.7% absolute. Finally, in section 5, detailed results and analysis for the combination of four

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. We would like to thank BBN Technologies and Petr Fousek for their assistance with this research.

individual system are presented. It is shown that incorporating morphological decomposition and BMMI into separate system branches improves the Cambridge Arabic STT system by up to 1.1% absolute WER.

2. BASELINE SYSTEMS

Graphemic and phonetic Arabic STT systems were built, and Multi-Layer Perceptron (MLP) acoustic features were used. These features are derived from an MLP and they directly encode high-level phonetic information. In the context of Arabic STT, these targets offer modelling advantages since they can be trained using graphemic and phonetic targets [3, 4, 5].

For graphemic systems, the most obvious MLP training targets are graphemic units. However, during training, a complete phonetic alignment of the training data is given which also allows the use of phonetic targets. Compared to a network trained on graphemic targets, a more powerful phonetic MLP can easily be obtained which codes high-level vowel information in its feature stream. Consequently, by integrating the MLP features within a tandem-connectionist framework [6] into an STT system, an MLP trained on phonetic targets can also be applied to a graphemic system. During the development of the graphemic system, the use of a graphemic and a phonetic target MLP was explored, and, within the framework of a development setup, the phonetic target MLP achieved an up to 0.4% better performance in terms of absolute WER. This confirms that the phonetic target MLP provides phonetic information in its feature stream which is otherwise not accessible to the graphemic system. Consequently, an MLP trained on phonetic targets was used for all phonetic and graphemic systems.

PLP-based and tandem PLP+MLP-based front-ends were built. The former used a 39-dimensional feature vector, and this was augmented by a 26-dimensional MLP feature vector to obtain the PLP+MLP front-end. The network layout and training are described in [7]. All systems were trained using 1538 hours of acoustic data. Cross-word decision-tree state-clustered triphones were built using MPE training, and gender-dependent models were constructed. First the PLP-based systems were trained. Then the tandem PLP+MLP feature systems were created by splicing MLP features into the PLP feature stream and applying the ‘fast system build’ described in [7]. The Language Models (LMs) were 4-grams trained on 1G words and built as in [8].

A multi-pass adaptation framework was used to evaluate the systems (see Fig.1). This is a three-stage process. The P1-stage is a fast decoding run with gender independent (GI) PLP graphemic models. The P2-stage uses speaker adapted gender dependent (GD) graphemic models based on the P1 supervision. The P2-stage generates trigram lattices which are expanded using a 4-gram language model and then rescored in the P3 stage. The P3-stage models are again GD models (both graphemic and phonetic, PLP and

PLP+MLP), adapted using 1-best CMLLR and lattice-MLLR as discussed in [1]. Confusion network decoding was then performed on this output and ROVER [9] was used for system combination. For PLP-system adaptation, full CMLLR and lattice-MLLR transforms were used. For the PLP+MLP-systems, block diagonal transforms were deployed, one block for the PLP features, and one for the MLP features.

For lattice generation by the P1+P2 stage, a graphemic PLP-based system was always used. Therefore the acoustic systems differ primarily at the P3 lattice rescoring stage (as discussed below). However, two different graphemic systems were used for P1+P2 decoding. Depending on the kind of acoustic models used for the P3 stage, either a word-based or a MADA-based system was used.

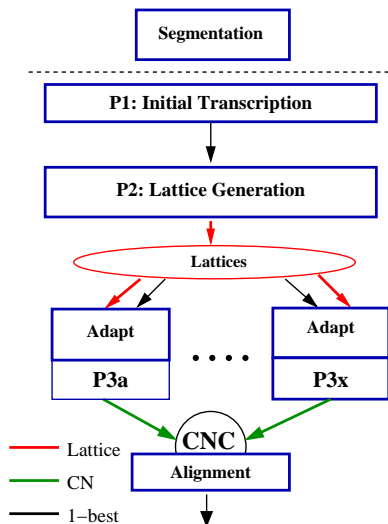


Fig. 1. The multi-pass and multi-branch framework.

3. MORPHOLOGICAL DECOMPOSITION

Arabic is a highly inflected, morphologically rich language, and in order to reduce OOV rates, software such as ‘Morphological Analysis and Disambiguation for Arabic’ (MADA) tools can be used to obtain a morphological decomposition of the vocabulary in both the training and test sets [10].¹ MADA implements a tokenisation and tagging stage, followed by a morphological disambiguation stage. The final output contains decomposed morpheme sequences. Graded levels of morphological detail are available, and the ‘D2’ configuration gives the best Statistical Machine Translation (SMT) and STT results [12, 8]. This configuration identifies five proclitics (l , b , k , w , f) and separates them from their associated word roots. For instance, the complex structure ‘ $lrjAl$ ’ (literally, ‘to a man’) is decomposed into the morpheme sequence ‘ $l + rjAl$ ’.

There is no bidirectional mapping between the MADA domain and the word domain. Apart from morphological analysis and decomposition, MADA also attempts to resolve ambiguities in the Arabic vocabulary by regenerating and sometimes normalising the lexical roots. The morpheme-to-word conversion process can be viewed

¹The MADA tools can be obtained from <http://www1.cs.columbia.edu/~rambow/software-downloads>. For another recent attempt to incorporate MADA processing into STT, see [11].

as an SMT task: the MADA domain is the source ‘language’; the word domain is the target ‘language’, and a linear alignment between the source and target token sequences can be obtained. As there are only many-to-one mappings from the source to the target, and no mappings to ‘NULL’, an N-gram SMT approach was used, as described in [8].

3.1. MADA Vowelisation

MADA can produce vowelised forms for input graphemic words. Therefore, two types of word and morpheme lists were created. The first contained the 350K most frequent words, and the 331K most frequent morphemes (determined using weighted combinations of all the acoustic training sources). The second type contained 260K word and 171K morpheme ‘phonetic’ subsets of the corresponding 350K word and 331K morpheme lists, derived using Buckwalter and MADA respectively.

Since the vowelisation rules applied by MADA are context-sensitive, the phonetic morpheme dictionary generation could not be based only on the corresponding morpheme list. As the context needed to be taken into account, the dictionary generation required the complete LM training data to be processed using MADA. The resulting MADA streams provided a one-to-one mapping between the graphemic and phonetic forms which were collected and merged to create a master dictionary. Based on this master dictionary, pronunciations could be obtained for 171K of the 331K morphemes. For the remaining 90K words and 160K morphemes which were not covered by Buckwalter or the morpheme master dictionary, pronunciations were obtained using the automatically derived rules described in [13].

The Out-Of-Vocabulary (OOV) rates for the three test sets used are given in Table 1. System performance was evaluated on three test sets dev07 (2.58 hours) dev08 (3.04 hours) and a set not used for development eval07 (2.00 hours)². All these test sets consist of both Broadcast News (BN) and Broadcast Conversation (BC) data.

| Wordlist | | Testset | | |
|----------|------|---------|--------|-------|
| Type | Size | dev07 | eval07 | dev08 |
| Word-pho | 260k | 2.68 | 3.39 | 2.03 |
| Word-gra | 350k | 1.19 | 1.26 | 1.14 |
| MADA-pho | 171k | 1.40 | 1.33 | 1.38 |
| MADA-gra | 331k | 0.79 | 0.54 | 0.63 |

Table 1. OOV rates for the word-based and MADA-based wordlists.

Like Buckwalter, MADA often generates several alternative vowelised forms for a given graphemic morpheme. However, in contrast to the word-based case (where the word vowelisation is obtained using Buckwalter), MADA ranks the possible vowelisations and outputs only the form with the highest score (1best). For the work described here, the phonetic morpheme dictionary was created by selecting the MADA 1best vowelisation for each lexical item. Pronunciation probabilities were obtained by aligning the acoustic training data using the MADA 1best dictionary.

In Table 2 (and henceforth), ‘G’ denotes a graphemic system, while ‘V’ (for Vowelised) denotes a phonetic system. If no further subscript is given, then a word-based MPE-trained system which uses PLP features is indicated. The use of MADA morphological decomposition, tandem PLP+MLP features, or BMMI training is indicated by the subscripts M_a , T_a , or B_m , respectively. The results

²The non-sequestered version of eval07 was used.

| System | Configuration | Testset | | |
|-----------------------|---------------|---------|--------|-------|
| | mada | dev07 | eval07 | dev08 |
| G | - | 13.2 | 14.1 | 14.9 |
| G_{Ma} | ✓ | 12.8 | 13.6 | 14.3 |
| V | - | 11.4 | 12.9 | 14.0 |
| V_{Ma} | ✓ | 11.0 | 12.4 | 13.5 |

Table 2. Contrast of a graphemic and a phonetic word-based system with their morpheme-based counterparts, WER in %.

in Table 2 contrast MADA-based graphemic and phonetic systems. The graphemic MADA decomposition gives WER gains of 0.4-0.6% absolute over the word-based baseline. For the phonetic MADA system, the gains over the baseline are similar (0.4-0.5% absolute).

4. BOOSTED MMI

Boosted MMI (BMMI) is a discriminative acoustic model training criterion [14, 15], and the objective function is:

$$\mathcal{F}_{\text{BMMI}}(\lambda) = \sum_r \log \frac{p(\mathbf{O}^{(r)} | \mathbf{w}_{ref}^{(r)}; \lambda)^\kappa P(\mathbf{w}_{ref}^{(r)})}{\sum_{\mathbf{w}'} p(\mathbf{O}^{(r)} | \mathbf{w}'; \lambda)^\kappa P(\mathbf{w}') e^{-\alpha \mathcal{A}(\mathbf{w}', \mathbf{w}_{ref}^{(r)})}} \quad (1)$$

where α is the boosting factor and $\mathcal{A}(\mathbf{w}, \mathbf{w}_{ref}^{(r)})$ is the phone accuracy. BMMI is effectively a modified MMI training scheme where the likelihoods of the competitor sentences are penalised by the boosting term $e^{-\alpha \mathcal{A}(\mathbf{w}, \mathbf{w}_{ref}^{(r)})}$, giving more weight to more confusable data.

BMMI can be compared to Minimum Phone Error (MPE) training, where

$$\mathcal{F}_{\text{MPE}}(\lambda) = \sum_r \frac{\sum_{\mathbf{w}} p(\mathbf{O}^{(r)} | \mathbf{w}; \lambda)^\kappa P(\mathbf{w}) \mathcal{A}(\mathbf{w}, \mathbf{w}_{ref}^{(r)})}{\sum_{\mathbf{w}'} p(\mathbf{O}^{(r)} | \mathbf{w}'; \lambda)^\kappa P(\mathbf{w}')} \quad (2)$$

is maximised [16]. In both cases, (1) and (2), the phone accuracy $\mathcal{A}(\mathbf{w}, \mathbf{w}_{ref}^{(r)})$ emphasises the more confusable data during the parameter optimisation. Though MPE and BMMI are quite different, the common use of the phone accuracy establishes an affinity between the methods [17]. Consequently, it is of interest to determine the extent to which the performance of systems trained using these criteria differ. Since system combination is normally used, the complementary of the different branches is often more important than the independent system performance.

Table 3 compares the performance of MPE with BMMI for the PLP front-end, word-based V and V_{Bm} systems, and the tandem PLP+MLP front-end, MADA-based V_{TaMa} , and V_{TaMaBm} systems. In both cases, Confusion Network Combination (CNC) WERs are also given. For the V and V_{Bm} systems, BMMI performs 0.2-0.5% better than MPE, and for the V_{TaMa} and V_{TaMaBm} systems improvements of 0.1-0.5% are observed. Interestingly, combining the corresponding MPE and BMMI systems gives similar gains: 0.2-0.4% for the $V \oplus V_{Bm}$ combination and 0.1-0.3% for the $V_{TaMa} \oplus V_{TaMaBm}$ combination. This suggests that, despite their aforementioned affinity, the MPE-trained models and BMMI-trained models exhibit a complementarity which can be exploited by system combination.

| System | Configuration | | | | Testset | | |
|---------------------------|------------------------------|-----|------|------|---------|--------|-------|
| | plp+mlp | wrd | mada | bmmi | dev07 | eval07 | dev08 |
| V | - | ✓ | - | - | 11.4 | 12.9 | 14.0 |
| V_{Bm} | - | ✓ | - | ✓ | 11.2 | 12.6 | 13.5 |
| CNC | $V \oplus V_{Bm}$ | | | | 10.9 | 12.2 | 13.3 |
| V_{TaMa} | ✓ | - | ✓ | - | 10.8 | 11.7 | 12.8 |
| V_{TaMaBm} | ✓ | - | ✓ | ✓ | 10.7 | 11.6 | 12.3 |
| CNC | $V_{TaMa} \oplus V_{TaMaBm}$ | | | | 10.4 | 11.5 | 12.2 |

Table 3. MPE/BMMI single branch and CNC contrast, WER in %.

5. SINGLE BEST AND COMBINATION RESULTS

5.1. The Single Best System

Table 4 shows the development of the single best system by incorporating MADA morphological decomposition, MLP features, and Boosted MMI into the phonetic P3 branch. The starting point are the phonetic, word-based, and MPE-trained V models, using a PLP front-end. Incorporating either tandem PLP+MLP features (V_{Ta}) or MADA morphological decomposition (V_{Ma}) results in gains of up to 0.5% in absolute WER (though MADA performs slightly better), and both methods act in a highly complementary way. Simultaneously applying both methods (V_{TaMa}) gives absolute WER reductions of up to 1.2%, which is more than the aggregated performance improvements of the individual measures. Finally, the best individual system is obtained by applying BMMI instead of MPE for the acoustic model training (V_{TaMaBm}). Switching to BMMI results in a further reduction of 0.1-0.5% in terms of absolute WER.

| System | Configuration | | | | Testset | | |
|---------------------------|---------------|-----|------|------|---------|--------|-------|
| | plp+mlp | wrd | mada | bmmi | dev07 | eval07 | dev08 |
| V | - | ✓ | - | - | 11.4 | 12.9 | 14.0 |
| V_{Ta} | ✓ | ✓ | - | - | 11.3 | 12.4 | 13.7 |
| V_{Ma} | - | - | ✓ | - | 11.0 | 12.4 | 13.5 |
| V_{TaMa} | ✓ | - | ✓ | - | 10.8 | 11.7 | 12.8 |
| V_{TaMaBm} | ✓ | - | ✓ | ✓ | 10.7 | 11.6 | 12.3 |

Table 4. Development of the single best branch, WER in %

In summary, the absolute overall reduction in WER by the joint use of tandem PLP-MLP features, MADA morphological decomposition, and BMMI ranges from 0.7% to 1.7% in absolute WER (6.1-12.1% relative).

5.2. System Combination

System combination was performed using ROVER, and the combination of four branches was investigated. For each individual system, a range of configurations is possible. Table 5 lists those systems whose combination produced the lowest WERs. The baselines were given by the graphemic G and the phonetic V systems which both use word-based MPE-trained models, with a PLP front-end. Based on these systems, tandem PLP+MLP features (Ta), MADA morphological decomposition (Ma), and BMMI (Bm) were introduced.

The combination baseline (RO_{base}) was obtained from four word-based and MPE-trained systems which are distinguished by being either graphemic or phonetic, and by applying either PLP or tandem PLP+MLP features. To investigate the impact of MADA

on the combination performance, the PLP-based graphemic G system and the tandem PLP+MLP-based phonetic V_{Ta} system were replaced by their MADA-based G_{Ma} and V_{TaMa} counterparts. As shown in Table 5, the resulting RO_{Ma} combination system gives a gain of 0.8-0.9% in absolute WER over the RO_{base} baseline. Replacing the phonetic word-based PLP feature V system with a corresponding more powerful MADA-based system did not help. The gains were smaller, confirming that it is generally more advantageous to combine good complementary systems rather than better systems that are less complementary.

| System | Configuration | | | | Testset | | |
|--------------|--|-----|------|------|---------|--------|-------|
| | plp+mlp | wrd | mada | bmmi | dev07 | eval07 | dev08 |
| G | - | ✓ | - | - | 13.2 | 14.1 | 14.9 |
| G_{Ta} | ✓ | ✓ | - | - | 12.6 | 13.4 | 14.4 |
| V | - | ✓ | - | - | 11.4 | 12.9 | 14.0 |
| V_{Ta} | ✓ | ✓ | - | - | 11.3 | 12.4 | 13.7 |
| G_{Ma} | - | - | ✓ | - | 12.8 | 13.6 | 14.3 |
| V_{Bm} | - | ✓ | - | ✓ | 11.2 | 12.6 | 13.5 |
| V_{TaMa} | ✓ | - | ✓ | - | 10.8 | 11.7 | 12.8 |
| V_{TaMaBm} | ✓ | - | ✓ | ✓ | 10.7 | 11.6 | 12.3 |
| RO_{base} | $G \oplus G_{Ta} \oplus V \oplus V_{Ta}$ | | | | 10.8 | 11.9 | 12.8 |
| RO_{Ma} | $G_{Ma} \oplus G_{Ta} \oplus V \oplus V_{TaMa}$ | | | | 10.0 | 11.0 | 11.9 |
| RO_{MaBm} | $G_{Ma} \oplus G_{Ta} \oplus V_{Bm} \oplus V_{TaMaBm}$ | | | | 9.9 | 10.8 | 11.7 |

Table 5. 4-way system combination (ROVER) results, WER in %

Next, the impact of BMMI was explored. To maintain good complementarity of the individual systems, the acoustic models used in two branches (those produced by the V and V_{TaMa} systems) were replaced by corresponding BMMI-trained models. The RO_{MaBm} results in Table 5 show that this gave additional gains of 0.1-0.2% absolute WER over the RO_{Ma} results.

Comparing the final RO_{MaBm} result with the best single-branch V_{TaMaBm} system, gains of 0.6-0.8% in absolute WER were obtained, confirming the potential of system combination. Finally, when comparing the new best RO_{MaBm} system with the RO_{base} system, gains of 0.9-1.1% absolute WER (8.3-9.2% relative) were observed which can be attributed to the use of both MADA morphological decomposition and BMMI training.

6. CONCLUSIONS

This paper has described recent improvements to the Cambridge Arabic STT system. It has been shown that the use of phonetic MLP training targets provides a convenient way to incorporate otherwise inaccessible phonetic information into graphemic systems. The use of MADA for morphological decomposition has been extended to phonetic systems, and the associated dictionary generation was discussed in detail. Further, boosted MMI training was examined within the scope both in single-branch systems and in system combination where it was found that MPE- and BMMI-trained models act in a complementary fashion. For a LVCSR state-of-the-art single-branch system, the combined use of MLP features, MADA morphological decomposition, and BMMI results in WER reductions of up to 12.1% relative, and, within a state-of-the-art 4-way LVCSR system combination framework, the combined use of MADA and BMMI gives WER gains of up to 9.2% relative.

7. REFERENCES

- [1] M.J.F. Gales, F. Diehl, C.K. Raut, M. Tomalin, P.C. Woodland, and K. Yu, "Development of a phonetic system for large vocabulary arabic speech recognition," in *Proc. of ASRU*, 2007.
- [2] L. Nguyen, T. Ng, K. Nguyen, R. Zbib, and J. Makhoul, "Lexical and phonetic modeling for arabic automatic speech recognition," in *Proc. of Interspeech*, 2009.
- [3] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of ICASSP*, 2007, pp. 757–760.
- [4] P. Fousek, L. Lamel, and J.-L. Gauvain, "On the use of MLP features for broadcast news transcription," in *Lecture Notes in Computer Science*. Springer Verlag, 2008, pp. 303–310.
- [5] J. Park, F. Diehl, M.J.F. Gales, M. Tomalin, and P.C. Woodland, "Efficient generation and use of MLP features for arabic speech recognition," in *Proc. of Interspeech*, 2009.
- [6] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. of ICASSP*, 2000.
- [7] J. Park, F. Diehl, M.J.F. Gales, M. Tomalin, and P.C. Woodland, "Training and adapting MLP features for Arabic speech recognition," in *Proc. of ICASSP*, 2009, pp. 4461–4464.
- [8] F. Diehl, M.J.F. Gales, M. Tomalin, and P.C. Woodland, "Morphological analysis and decomposition for arabic speech-to-text systems," in *Proc. of InterSpeech*, 2009.
- [9] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," in *Proc. of ASRU*, 2006.
- [10] N. Habash and O. Rambow, "Arabic tokenisation, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proc. of ACL*, 2005.
- [11] D. Vergyri, A. Mandal, W. Wang, A. Stolcke, J. Zheng, M. Gra-ciarena, D. Rybach, C. Gollan, Schlüter R., K. Kirchhoff, A. Fari, and N. Morgan, "Development of the SRI/nightingale arabic ASR system," in *Proc. of Interspeech*, 2008.
- [12] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proc. of HLT-NAACL*, 2006.
- [13] F. Diehl, M.J.F. Gales, M. Tomalin, and P.C. Woodland, "Phonetic pronunciations for arabic speech-to-text systems," in *Proc. of ICASSP*, 2008.
- [14] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. of ICASSP*, 2008.
- [15] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," in *Proc. of Interspeech*, 2008.
- [16] D. Povey, "Discriminative training for large vocabulary speech recognition," *Ph.D Thesis, University of Cambridge*, 2004.
- [17] E. McDermott, S. Watanabe, and A. Nakamura, "Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training," in *Proc. of Interspeech*, 2009.