# TOWARDS ACOUSTIC MODEL UNIFICATION ACROSS DIALECTS

*Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro Moreno, Austin Waters*

Google Inc.

{mgelfeky, mbastani, xavvelez, pedro, austinwaters}@google.com

## ABSTRACT

Acoustic model performance typically decreases when evaluated on a dialectal variation of the same language that was not used during training. Similarly, models simultaneously trained on a group of dialects tend to underperform dialect-specific models. In this paper, we report on our efforts towards building a unified acoustic model that can serve a multi-dialectal language. Two techniques are presented: Distillation and MultiTask Learning (MTL). In Distillation, we use an ensemble of dialect-specific acoustic models and distill its knowledge in a single model. In MTL, we utilize multitask learning to train a unified acoustic model that learns to distinguish dialects as a side task. We show that both techniques are superior to the jointly-trained model that is trained on all dialectal data, reducing word error rates by $4.2\%$ and $0.6\%$, respectively. While achieving this improvement, neither technique degrades the performance of the dialect-specific models by more than $3.4\%$.

***Index Terms—*** multi-dialectal languages, acoustic modeling, knowledge distillation, multitask learning

## 1. INTRODUCTION

Dialects are defined as variations of the same language, specific to geographical regions or social groups. For example, Mexican, Argentine and Castilian are all different Spanish dialects. Similarly, Egyptian, Gulf, Levantine and Maghrebi are the main four Arabic dialects. Although dialects of the same language share many similarities, they are often differentiated at several linguistic levels; amongst others: phonological, grammatical, orthographic (e.g., "center" vs. "centre") and very often different vocabularies. Multi-dialectal languages can also be characterized by the level of intelligibility between its speakers. While dialectal Spanish speakers can generally understand each other without much difficulty, this is not the case for Arabic [1]. As a result, the computational tools trained or tuned for one specific dialect break or underperform when tested on another dialect from the same dialectal group (e.g., [2]). Similarly, global tools simultaneously trained on many dialects fail to generalize well for any of them, as will be shown later in the results.

Inevitably, multi-dialectal languages pose a challenge to speech recognition systems, that is, whether to develop a single unified recognizer that understands all dialects, or build a different recognizer per dialect. So far, the strategy at Google has been to build dialect-specific recognizers. This decision was based on linguistic facts as well as rigorous cross-dialect experimental analysis (e.g., [2]). In an application like VoiceSearch, the issue becomes how to choose a recognizer to decode an arbitrary spoken query. In the past, queries have been directed to a dialectal recognizer based on the user's language / country preferences; whereas, presently the recognizer is selected based on location information extracted from the query's IP address [3]. For example, voice queries that originate from Egypt are directed to the Egyptian Arabic recognizer.

In this paper, our objective is to develop a single unified dialect-independent recognizer. Having as few recognizers as possible is clearly a win from the deployment point of view, especially when supporting a large number of languages (Google, for example, supports recognition of almost 60 languages / dialects). Each recognizer has a large memory footprint that stems from the number of parameters its acoustic model needs.

A simple way to develop a single unified acoustic model is to jointly train the model using all available dialectal data. This becomes our baseline model that we are trying to outperform. We are proposing two techniques to develop such single unified model superior to the *jointly-trained* one. In Section 2, we propose the use of knowledge distillation, namely DKD (Dialectal Knowledge Distillation). In Section 3, we propose the use of multitask learning, namely DMTL (Dialectal MultiTask Learning). In both sections, we start by briefly introducing the concept, and then proceed to describe our proposed technique. Evaluation results for both techniques are presented in Section 4. We use Arabic as our experimental language due to its dialect-unintelligibility. Finally, Section 5 concludes the paper and discusses a number of possible future research avenues.

To our knowledge, only two recent works have attempted to develop such single unified dialect-independent recognizer. In [4], the authors have proposed a method to combine / select the best decoded hypothesis from the pool of dialectal recognizers. However, their method would require that each utterance gets decoded by all the dialectal recognizers, and hence does not yield to any memory reduction. In [5], the authors have developed what they called a *multi-accent* acoustic model for English British and Indian accents using adaptation. This work is more related to our MTL work and hence will be noted in Section 5.

## 2. KNOWLEDGE DISTILLATION

In a classification problem, knowledge distillation (KD) refers to training a *student* model on the output class probabilities of a *teacher* model [6]. KD's goal is to transfer the knowledge of a potentially very large teacher model to a smaller student model, which is more suitable for deployment. In speech recognition, knowledge distillation from an ensemble of acoustic models built for a single language/dialect has proven to be efficient and to produce better results than the individual models [7]. Chebotar and Waters use an ensemble of acoustic models that are trained with different architectures (e.g., Long Short-Term Memory (LSTM) [8, 9] and Convolutional LSTM Deep Neural Networks (CLDNN) [10]), and are trained with either cross-entropy (CE) or state-level minimum Bayes risk (sMBR) [11] criterion. In the next section, we propose how to use the same idea to create a student acoustic model that unifies dialectal acoustic models.

### 2.1. DKD: Dialectal Knowledge Distillation

The intuition behind using knowledge distillation to build a unified acoustic model for a multi-dialectal language stems from the fact that dialects that belong to the same language share a significant number of acoustic features. One can think of the ensemble of dialect-specific acoustic models as the teacher model that models the entire language. Using KD, the knowledge of this teacher model can be distilled into a student model that ideally would benefit from the similarities in the acoustic features of the language's dialects. For example, starting from an ensemble of Egyptian, Levantine, and Gulf Arabic acoustic models that were not necessarily built using the same architecture, we can use the same knowledge distillation technique used in [7] to build a unified Arabic acoustic model. Each dialect-specific acoustic model is trained solely with the corresponding dialectal data. However, to create an ensemble of these dialect-specific models, it is required that either all of them have the same context-dependent (CD) state inventory output, or there exists a mapping from each model's output to a unified one. We elect the former option and satisfy this requirement by building the CD tree in advance using all dialectal data. The resulting tree represents all the dialects with unified CD states. Our tree-clustering algorithm is the standard one of Young et al. [12]. Consequently, the ensemble dialect-specific models are trained to produce that unified CD states. Then, a linear combination function (e.g., weighted average) is used to combine the ensemble models output, which provides the training data for the student model. Our proposed technique is outlined in Algorithm 1, and illustrated in Fig. 1.

**Input:** Multi-dialectal training data (utterances and their corresponding correct transcripts).

1: Build a unified CD tree using all training data.
2: For each dialect, train a *dialect-specific* acoustic model using the dialect-specific training data, which outputs the posterior probabilities on the unified CD states.
3: Determine the optimal weights to combine the frame-level predictions of all *dialect-specific* models (created in the previous step). The optimal weights are found by performing a grid search over all possible weight combinations and choosing the one that leads to the best performance (word error rate) of the ensemble on a test set comprising the union of all dialectal test sets.
4: Train a *student* model using the ensemble of *dialect-specific* models as the teacher model.

**Output:** The *student* model, now representing the unified *dialect-independent* model for the language in hand.

**Algorithm 1:** Dialectal Knowledge Distillation

## 3. MTL: MULTITASK LEARNING

State-of-the-art acoustic models for automatic speech recognition (ASR) are typically based on deep neural networks (DNN) trained to predict the posterior probabilities of a set of context-dependent (CD) states [13]. Multitask learning (MTL) paradigm proposed in [14] was recently applied to acoustic modeling [15, 16, 17]. In MTL, multiple related tasks are jointly learned, such as for example predicting CD states and context-dependent graphemes simultaneously [16]. It was observed that when both the input features and the hidden units are shared, MTL learning may improve

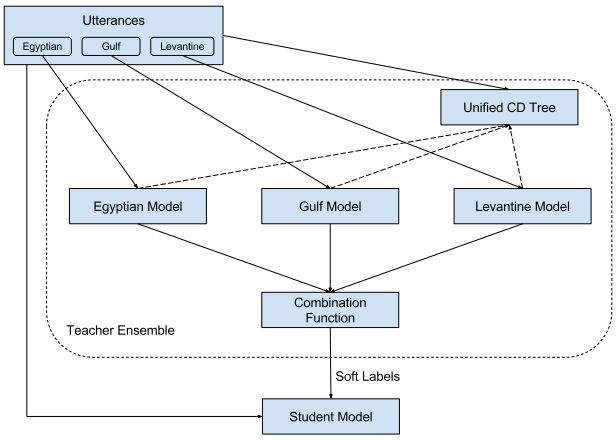


**Fig. 1**: DKD Framework

model's generalization ability as learning one task may help learning other tasks [14]. In acoustic modeling MTL scenario, a DNN is trained to predict CD states as the primary task and some other feature as a secondary task. Examples of the secondary feature to be learned include phone-labels, state-contexts, phone-contexts [15]; monophone-targets [17]; and trigrapheme-classes [16]. Usually, the secondary outputs are ignored during recognition and only the primary CD states posteriors are used to recognize the utterance in hand.

### 3.1. DMTL: Dialectal MTL

We propose to use MTL for building a unified acoustic model for a multi-dialectal language in a similar way to the one described above. In DMTL, we train a single DNN that predicts the CD states as the primary task and the dialect as the secondary task. This is achieved by adding a secondary output layer to the DNN while sharing the input and all the hidden layers. This secondary output layer is a softmax layer whose target output is a binary vector that represents the dialect of the input utterance. For a language with $n$ dialects, the secondary target output is a binary vector of size $n$ with zeros in all but the $i$-th position to indicate the $i$-th dialect. Fig. 2 depicts how a DMTL network looks like. The training process of DMTL-DNN is identical to that of a normal DNN [13] using cross-entropy, while also taking into account the gradients of the secondary targets.

At recognition time, the secondary targets are ignored and only the CD states primary targets are used to recognize the utterance. The intuition is that by letting the DMTL network learn about the dialect, it can serve as a *dialect-independent* single acoustic model for the language in hand.

## 4. EXPERIMENTAL RESULTS

As mentioned in Section 1, Arabic is the language of choice for our experiments, and three dialects of choice are Egyptian, Gulf and Levantine. Each dialect has a training corpus of around 3M anonymized user utterances (approximately 2,000 hours), and a testing set of 25K anonymized and manually transcribed utterances (approximately 12.5 hours). Manual transcribers were asked to follow specific guidelines that take into account the spoken form of the Arabic dialects, and the plethora of colloquial words in them. The test

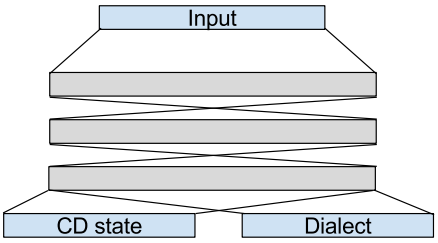


**Fig. 2**: DMTL Network

sets are 50% balanced between voice-search (VS) and dictation (DT) logs. Hence, there are six test sets: two test sets per dialect of dictation and voice-search. For evaluation purposes, we measure word error rates (WERs) on the six test sets. We use a similar CLDNN architecture to [10] with CE training for all evaluated models. Specifically, we use one convolutional layer, a dimensionality-reduction layer, three LSTM layers, and finally two DNN layers, for more details about each layer, see [10].

To evaluate our proposed techniques, we compare the performance (WER) of the *DKD* and the *DMTL* CLDNN models against two groups of models:

- Three *dialect-specific* CLDNN models, each trained solely using the dialect-specific data, each with its own CD states inventory.
- A *jointly-trained dialect-independent* CLDNN model trained with all available training data. In other words, this is the model that ignores the multi-dialectal nature of Arabic.

It is important to note that the second model above is the baseline model we are most interested in comparing against in order to achieve our objective: a single *dialect-independent* model that outperforms the simple *jointly-trained* model. We are including the *dialect-specific* results not only for completeness, but also for pointing out our direction for future work (see Section 5).

Before evaluating our proposed models, let us compare the performance of the baseline models. As expected, Table 1 shows that the *jointly-trained dialect-independent* model generally underperforms the *dialect-specific* models. The quality losses on all, but the Egyptain, test sets are as expected, but we could not explain why Egyptian is an exception.

### 4.1. DKD Results

In addition to the two groups of models mentioned above, and in order to be able to evaluate the impact of unifying the CD states inventory, we compare the performance (WER) against a third group:

- Three *unified-cd-states dialect-specific* CLDNN models, each trained solely using the dialect-specific data, yet using the unified CD states. These models are the ones used to create the *teacher* ensemble.

Table 1 shows that the *unified-cd-states dialect-specific* models underperform the *cd-states-specific dialect-specific* models on almost all test sets. This is likely due to the latter being able to benefit from a CD states inventory optimized for the specific dialect of each model.

Tables 1 and 2a show clearly that the *DKD* model outperforms the *jointly-trained* model in all of the six test sets, with relative WER improvements ranging from $1.9\%$ to $6.7\%$. With respect to the *dialect-specific* models, although the *DKD* model does not outperform all individual *dialect-specific* models, on average over all test sets, it improves relative WER by $0.4\%$. In other words, the gains on the Egyptian test sets wash out the losses on the other test sets. Therefore, our *DKD* model is achieving our objective of outperforming the *jointly-trained* model without degrading the performance when compared to the *dialect-specific* models. Lastly, the *DKD* model has an average quality gain of $7.1\%$ over the *unified-cd-states dialect-specific* model.

### 4.2. DMTL Results

We evaluate our *DMTL* CLDNN model against the two aforementioned baseline models: the *dialect-specific* and the *jointly-trained dialect-independent* models.

Tables 1 and 2b show that the *DMTL* model slightly outperforms the *jointly-trained* model in all the six test sets, with average relative WER improvement of $0.6\%$. While this might seem small, the WER reductions are statistically significant in the context of our typical 12.5 hour test sets. In addition, given the scale of Google's voice search traffic, our experience has been that even such small WER reductions translate into significant increases in user satisfaction metrics such as click through rates, lower retries, and lower rejections, etc. Therefore, this result is achieving our objective: a single *dialect-independent* model for Arabic that performs better than the *jointly-trained* model. Lastly, similar to the DKD results, the *DMTL* model does not outperform all individual *dialect-specific* models, except for the Egyptian test sets. However, let us reiterate that comparing our models to the *dialect-specific* models is not the main objective of this paper.

Comparing the WER performance of DKD against DMTL, it is clear that the former technique is superior. Nevertheless, we just experimented with the simplest idea of employing MTL for dialects, and yet got promising results. Among the two, MTL is very extensible. For example, MTL can work readily with any acoustic model training method. In contrast, using KD with the most-recent connectionist temporal classification (CTC) [18, 19] training method is not trivial since the underlying acoustic models are not guaranteed to have the same time-alignments (phone spike outputs). At the end of the next section, we provide some potential future work to improve our MTL technique.

## 5. CONCLUSION AND FUTURE WORK

With an objective of developing a single unified acoustic model for a multi-dialectal language, we presented two techniques: *DKD* and *DMTL*. In DKD, we used the idea of knowledge distillation to train a single deployable acoustic model on the outputs of an ensemble of dialect-specific acoustic models. In our second technique, DMTL, we utilized the idea of multitask learning to create a dialect aware acoustic model. Our experimental results, on three dialects of Arabic language, clearly showed that both of these ideas outperform the simple acoustic model that is trained on all dialectal data. In conclusion, we showed that both of these techniques have the potential to be adopted in large scale speech recognition systems. The techniques can reduce the overall speech recognition system footprint across dialects without reducing the overall quality of the recognizer system.

The future challenge is to devise a unification technique that on each dialect performs better than the dialect-specific acoustic model. The MTL technique has potential to be extensible and can be modified in various ways to address this challenge. For example, by

| Absolute WER | Test set | Dialect Specific | Unified CD States | Jointly Trained | DKD | DMTL |
|---|---|---|---|---|---|---|
| Egyptian | DT | 35.7 | 37.0 | 34.4 | 33.6 | 34.5 |
| | VS | 31.8 | 32.7 | 30.7 | 29.1 | 30.2 |
| Gulf | DT | 29.3 | 29.7 | 30.9 | 30.3 | 30.9 |
| | VS | 31.8 | 31.6 | 33.9 | 32.5 | 33.6 |
| Levant | DT | 23.7 | 27.2 | 26.2 | 24.9 | 25.9 |
| | VS | 20.7 | 25.9 | 22.5 | 21.0 | 22.4 |

**Table 1**: Absolute word error rates (WER) of DKD and DMTL vs. the base models

| Relative WER (%) | Test set | Dialect Specific | Unified CD States | Jointly Trained |
|---|---|---|---|---|
| Egyptian | DT | -5.9 | -9.2 | -2.3 |
| | VS | -8.5 | -11.0 | -5.2 |
| Gulf | DT | +3.4 | +2.0 | -1.9 |
| | VS | +2.2 | +2.8 | -4.1 |
| Levant | DT | +5.1 | -8.5 | -5.0 |
| | VS | +1.4 | -18.9 | -6.7 |

(a) DKD

| Relative WER (%) | Test set | Dialect Specific | Jointly Trained |
|---|---|---|---|
| Egyptian | DT | -3.4 | +0.3 |
| | VS | -5.0 | -1.6 |
| Gulf | DT | +5.5 | 0.0 |
| | VS | +5.7 | -0.9 |
| Levant | DT | +9.3 | -1.1 |
| | VS | +8.2 | -0.4 |

(b) DMTL

**Table 2**: Quality gain(-)/loss(+) of DKD and DMTL vs. the base models

further analysis of the secondary output layer impact on the network learning, we can experiment with, e.g., adding more layers in the secondary branch of the neural network, or using adaptation layers [5]. Another idea is to use domain-adversarial [20] learning for the DMTL network. That is, rather than the normal backpropagation on both output layers, the secondary output layer is trained to maximize accuracy while the shared hidden layers are trained adversarially to minimize accuracy.

## 6. ACKNOWLEDGMENT

The authors would like to thank Eugene Weinstein and Olivier Siohan for their valuable comments and discussion.

## 7. REFERENCES

[1] Margo E. Wilson, "Arabic speakers: Language and culture, here and abroad," *Topics in Language Disorders*, vol. 16, no. 4, 1996.

[2] F. Biadsy, P.J. Moreno, and M. Jansche, "Google's cross-dialect arabic voice search," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[3] M. Elfeky, P. Moreno, and V. Soto, "Multi-dialectical languages effect on speech recognition: Too much choice can hurt," in *International Conference on Natural Language and Speech Processing (ICNLSP)*, 2015.

[4] V. Soto, P. Moreno, and M. Elfeky, "Selection and combination of hypotheses
for dialectal speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[5] Y. Huang, D. Yu, C. Liu, and Y. Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Interspeech*, 2014.

[6] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computing Research Repository (CoRR)*, vol. abs/1503.02531, 2015.

[7] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Interspeech*, 2016.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[9] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014.

[10] T. N. Sainath, O. Vinyals, A. W. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[11] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[12] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA Human Language Technology Workshop*, 1994.

[13] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[14] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[15] M. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[16] D. Chen, B. Mak, C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[17] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[19] A. Senior, H. Sak, F. de Chaumont Quitry, T. N. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *Jouranl of Machine Learning Research*, vol. 17, pp. 1–35, 2016.