Investigating data sharing in speech recognition for an underresourced language: the case of algerian dialect

Mohamed Amine Menacer, Kamel Smaïli

▶ **To cite this version:**

Mohamed Amine Menacer, Kamel Smaïli. Investigating data sharing in speech recognition for an underresourced language: the case of algerian dialect. 7th International Conference on Natural Language Processing - NATP 2021, Mar 2021, Vienna, Austria. ⟨hal-03137048⟩

HAL Id: hal-03137048

https://hal.science/hal-03137048v1

Submitted on 10 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

INVESTIGATING DATA SHARING IN SPEECH RECOGNITION FOR AN UNDER-RESOURCED LANGUAGE: THE CASE OF ALGERIAN DIALECT

Mohamed Amine Menacer[1] and Kamel Smaïli[1]

[1]Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
mohamed-amine.menacer@loria.fr
kamel.smaili@loria.fr

ABSTRACT

The Arabic language has many varieties, including its standard form, Modern Standard Arabic (MSA), and its spoken forms, namely the dialects. Those dialects are representative examples of under-resourced languages for which automatic speech recognition is considered as an unresolved issue. To address this issue, we recorded several hours of spoken Algerian dialect and used them to train a baseline model. This model was boosted afterwards by taking advantage of other languages that impact this dialect by integrating their data in one large corpus and by investigating three approaches: multilingual training, multitask learning and transfer learning. The best performance was achieved using a limited and balanced amount of acoustic data from each additional language, as compared to the data size of the studied dialect. This approach led to an improvement of 3.8% in terms of word error rate in comparison to the baseline system trained only on the dialect data.

KEYWORDS

Automatic speech recognition, Algerian dialect, MSA, multilingual training, multitask learning, transfer learning.

1. INTRODUCTION

Arabic language comprises thirty modern varieties[1], including its standard form, Modern Standard Arabic (MSA), which is derived from Classical Arabic. MSA is a simplified version of the Classical Arabic (a literary form) with a modernized vocabulary. It is the official form used in the newspapers and in the formal communications. The other Arabic language varieties, referred as dialects, come from historical interactions between classical Arabic and languages of the regional cultures and from the linguistic influence due to colonization. They are used in the Arab world in informal conversational context and in the daily communication.

In many Natural Language Processing (NLP) applications, the bulk of works proposed in the literature is intended for MSA, less works are dedicated to Arabic dialects. For a long time, the NLP community was not interested by Arabic dialects, but nowadays a craze for these dialects has been observed. In fact, there are several reasons for that: the Arabic dialects constitute the daily language of communication in Arab world, they are under-resourced languages, there is no standardization for writing them, some of them are very different from MSA, they often are code-switched, etc. All these features make them challenging in point of view of NLP. In this article,

[1] Source: ISO 639-3 ara documentation

we focus on an Algerian Arabic dialect, the one used in Algiers and its periphery, for which we propose an Automatic Speech Recognition (ASR) system.

NLP for under-resourced languages, such as Arabic dialects, requires more sophisticated techniques that go far beyond the basic re-training of the models dedicated to well-resourced languages. The approaches that have been proposed so far to recognize under-resourced languages focused mainly on two aspects: proposing, on one hand, data collection methodologies and introduce, on the other hand, advanced training techniques to cope with the lack of data. To develop an ASR system for an under-resourced language, one needs firstly to collect the necessary data for its different components. Works on data collection are carried out via crowdsourcing [1] or via exploring data for which information are shared between languages [2, 3].

For the acoustic data, it is often difficult to obtain spoken transcribed resources for under-resourced languages. One can achieve this by transcribing existing audio resources [4] or by recording speech from existing textual data [5]. Concerning textual data, the easy way to collect them is to investigate web content [6].

Moreover, a pronunciation dictionary must be created; the grapheme-based approach is the simple way to produce it. One considers for Arabic that the pronunciation of each word is simply its grapheme decomposition, and therefore, graphemes represent the basic units for the Acoustic Model (AM) [3, 7]. Other approaches are used to convert graphemes to phonemes such as those based on statistical machine translation [8, 9] or on linguistic rules [10, 11].

Since the data collection for under-resourced languages is time consuming, unsupervised or semi-supervised approaches are pretty adequate in this context. One underlying technique that can be used when only a small amount of transcribed data is available is to develop a baseline ASR system and use afterwards this system to transcribe a large amount of data. These new transcribed data can be used to fine tune the baseline system and improve the speech recognition performance [12]. Another interesting approach is to take advantage from other languages. The idea is to develop a multilingual model that combine information from several languages that share words [2, 13].

For the Algerian dialect, there is no transcribed data for training the acoustic model. To handle this issue, we propose to record a small spoken corpus for developing a baseline system and then, to improve it by taking advantages from the speech data of other languages that impact the Algerian dialect.

2. ISSUES FOR DEVELOPING AN ASR SYSTEM FOR AN ALGERIAN DIALECT

The vocabulary used in the Algerian dialect comes from the historical interaction between multiple languages, namely MSA, French, Turkish and Berber. Words from these languages could be employed without any modification, or they could be altered to produce new words. This fact leads to a new language variety that is different from the MSA, and that can be defined as a mixture of several languages.

Because of the borrowed words, the phonetic system of the Algerian dialect is a mixture of Arabic phonemes and others especially used in the French language. This leads to an exhaustive list of 47 phonemes (34 Arabic phonemes plus 13 French phonemes). Consequently, to correctly recognize the Algerian dialect, the first issue that we need to handle is to train an acoustic model that recognizes all these phonemes.

The Algerian Arabic dialect is mainly spoken, that means that the way of writing is free. People could use Latin or Arabic script or mix all the foregoing in the same sentence to convey their ideas. Some examples of the writing system extracted from social networks are illustrated in Table 1.

Table 1. Examples of some writhing possibilities in the Algerian dialect.

Arabic script	الله يخليك عندي مشكل في ترتيب لفاليز دياولي كاش فكرة
Latin script	allah yekhalik aandi mochkil fitartib les valises dyawli kache fekra
Mix script	الله يخليك عندي problème في ترتيب les valises دياولي كاش idée
Translation	Please, I have a problem of organizing my suitcases, any idea!

In the following sections, the techniques used to model the acoustic and the language aspects for the Algerian dialect are set forth.

3. LANGUAGE MODELLING

Algerian dialect is mainly spoken without any conventional writing rules. Consequently, it is difficult to find well-formed text written in dialect. One way to deal with this issue is to retrieve textual data from social networks. In our previous works, two corpora containing Algerian dialects were constituted: PADIC [14, 15] and CALYOU [6] corpora.

- **PADIC** is a collection of 6400 Modern Arabic sentences with their translations in several Arabic dialects (Two from Algeria, Tunisian, Moroccan, Palestinian and Syrian). This corpus was developed manually by translating Arabic conversational sentences into the different dialect variants.
- **CALYOU** is a large corpus collected from comments of Algerian videos on YouTube. It contains 1.4M dialect sentences written in Arabic and Latin scripts.

While the writing system in PADIC corpus is standardized (by adopting some rules and by using Arabic characters extended with (پ/p/, ڤ/v/, ق/g/) for non-letters sounds), sentences in CALYOU corpus are not normalized, since it is a collection of comments extracted from social network, where the way of writing is free. For this reason, we carried out a pre-processing to normalize the data of CALYOU, it consists of:

- Removing all the sentences written or containing Latin script.
- All the homophones that have the same meaning are replaced by the most frequent spelling by using a lexicon proposed in [16]. Some examples are given in Table 2.

After having processed the CALYOU corpus, the total number of sentences is reduced to 650K.

Table 2. Examples of some homophones that have the same meaning.

Homophones	**Replaced by**	**Translation**
فيلم – فليم - فلم	فيلم	Film
منفقين - منافقين	منافقين	Hypocrites
خااوة – خوة – خوا – خاوى – خاوة – خاوا	خاوة	Brothers

The training of the Language Model (LM) for the Algerian dialect is not restricted on the two corpora PADIC and CALYOU, we also take advantage from MSA data. Since the amount of the different textual data is unbalanced, the LM, we propose, is a linear combination of four bigram models. Two of them are trained on MSA textual data: the MSA version of Gigaword (1 billion of word occurrences) and the transcripts of the MSA speech data used to train the acoustic models (315k words). The two others are trained on dialectal data: PADIC and CALYOU. The weights of the linear interpolation are estimated on a development corpus composed by a mixture of MSA

and dialect data. The resulting weights for each corpus are 0.48 for CALYOU, 0.22 for MSA Gigaword, 0.11 for PADIC and 0.19 for the transcripts of the MSA speech data.

4. PRONUNCIATION MODELLING

The lexicon is composed by the union of the most frequent words extracted from each dataset used for training the language model. For each word in the lexicon, one needs to have all its pronunciation variants. The issue is how to produce all possible pronunciation variants for Arabic words, among them a subset of dialect words, knowing that Arabic texts are written without any diacritic.

Since linguistic resources are available for MSA, we used an external lexicon [17] as a lookup table from which the pronunciations of the MSA words are extracted and inserted into the pronunciation lexicon of our ASR system. Unfortunately, we do not have the equivalent for the Algerian dialect. For this, we adopted a G2P approach to produce pronunciation variants for dialectal words. We adapted to our purpose the approach proposed in [10]. The conversion G2P process is based on two stages:

- Restore diacritics using a statistical approach. This issue is considered as a machine translation problem where the source language is a set of undiacritized texts and the target one is a set of diacritized texts. A Statistical Machine Translation (SMT) system was trained by using existing tools on a parallel corpus of undiacritized and diacritized Algerian dialect texts. Since this parallel corpus was built manually and the task of vocalization is time consuming, this corpus contains only 4k sentences. This approach led to a precision of 98% at the character level and 96% at the word level.
- Use a set of hand-crafted rules to produce the phonetic representation of the dialectal words. For further details about these rules, the reader is directed to the work [10].

5. ACOUSTIC MODELLING

The main challenge that we are facing is to get a spoken transcribed corpus for the Algerian dialect. Because recording is a costly task, we selected only 4.6k dialect sentences extracted from PADIC and CALYOU and we asked native Algerian speakers to record this small corpus. The selection is carried out in such a way that the length of the sentences fluctuates between 3 and 20 words with an average duration of 4.5 seconds.

Seven Algerian native speakers recorded, in a quiet room and using a professional unidirectional microphone, the selected corpus. Two of them are female and five are male.

The resulted corpus contains 6 hours of speech sampled at 16 kHz. This dataset, named ADIC (Algerian DIalect Corpus) is split into three parts as it is shown in Table 3. The speakers of the Test data are different from the ones of the Train and the Dev data.

Table 3. Some characteristics of ADIC.

Subset	Duration	Speakers
Train	240 min	4
Dev	40 min	
Test	75 min	3

5.1. Learning by using a TDNN architecture

We propose to use an acoustic model based on the time delay neural network (TDNN) architecture [18] as described in Table 4. TDNN is a kind of feed-forward neural network used to better handle the context information of speech signal through a carefully designed hierarchical structure [19].

It is based on the use of context windows where the input layer processes acoustic features with narrow contexts while wider contexts are processed by the deeper layers.

Each deep layer receives several outputs that are spliced from the previous layer. The first layer receives a concatenation of 5 acoustic features corresponding to the features from $t-2$ to $t+2$ (see line 2 of Table 4). In layer 2, we splice together the input at the current frame minus 1 until the current frame plus 2. This means that the second layer will capture implicitly a larger context of the acoustic features from $t-3$ to $t+4$.

In this case, one can understand that the number of parameters to train is huge. To deal with this issue, we adopt the method proposed in [18] to sub-sample the TDNN network. In this approach, instead of splicing all the frames, only two frames are gathered corresponding to the first and the last frame of the original method. For instance, the notation $\{-1,+2\}$ in the third line of Table 4, means that only the two outputs -1 and $+2$ are spliced. At the end, the output of the last layer handles implicitly the context of $[t-16, t+11]$ for each acoustic parameter at t timestamp.

Table 4. Context specification for each layer of the TDNN model.

Layer	**1**	**2**	**3**	**4**	**5**	**6**
Input context	$[-2,+2]$	$[-1,+2]$	$[-3,+3]$	$[-3,+2]$	$[-7,+2]$	$\{0\}$
Input context with sub-sampling	$\{-2,+2\}$	$\{-1,+2\}$	$\{-3,+3\}$	$\{-3,+2\}$	$\{-7,+2\}$	$\{0\}$

The training of the TDNN model is based on sMBR sequence-discriminative criterion [20] and the parameters are estimated with the stochastic gradient descent algorithm.

Since ADIC is considered small for training the TDNN model, our idea is to benefit of other languages that impact the dialect (MSA and French) and to transfer the acquired knowledge to the ASR of the dialect. To do so, we proposed three different approaches depending on how the MSA and French acoustic data are integrated into the training process of the acoustic model of the Algerian dialect. These approaches are the multilingual training, the multitask learning and the transfer learning (see Figure 1).

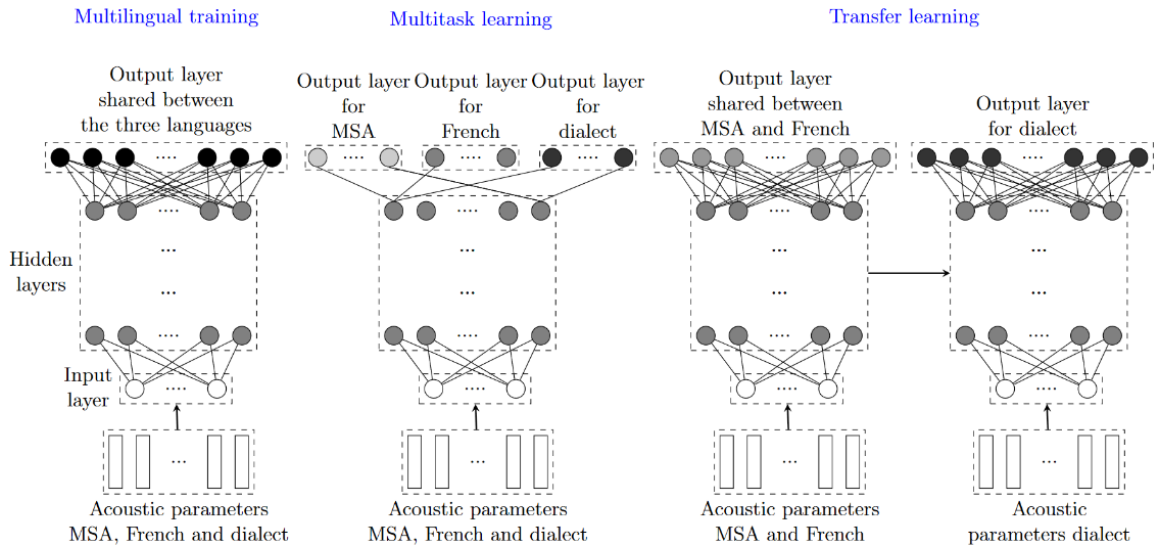

Figure 1. Proposed training techniques for the dialect acoustic modelling.

5.2. Multilingual training

In this approach we merged all the acoustic data of the three languages to construct a larger corpus. We then used it to train a TDNN without any distinction between the three languages. In this case, all the layers of the neural network are shared between the languages. Two questions were raised before the implementation of this solution:

- Knowing that MSA and French languages share some phonemes (e.g. /k/, /z/, etc.), how to find the best phonetic representation since the output layer that predicts triphones is shared between the three languages?
- How to optimize the necessary amount of speech data of each language (MSA and French) to make the contribution of each of them more effective on the performance of the ASR system of the Algerian dialect.

Concerning the first question, the integration of MSA and French data was carried out according to the two following approaches:

- **Union of phonemes** we simply take the union of the French and the MSA phonemes lists. This led to a set of 65 phonemes (34 MSA and 31 French).
- **Shared phonemes** the shared phonemes set is produced by keeping only one instance for each common phoneme. This led to a set of 47 phonemes (17 phonemes are common between MSA and French).

Concerning the optimization of the amount of data of each language, we decided to increase the training part of ADIC gradually by few hours of each language(MSA and French) until reaching a total of 44 hours and then we select the combination that performs better on the Dev part of ADIC.

Figure 2 indicates the evolution of the Word Error Rate (WER) while adding at each step 4 hours of French data. The number above each curve represents the amount of MSA data (in hours). The blue and the black plots represent respectively the evolution of the WER when using the union of phonemes and when using shared phonemes. The WER in the baseline system (without adding MSA nor French data) is 30.05%. The best results (a WER of 28%) is the one got by adding 12 hours of MSA data and 12 hours of French data (see the curve (d)).

The experiments show that when increasing considerably the amount of MSA and French data (more than 12 hours), the results decrease. This last remark was expected, but we learned from these experiments the exact amount of the data necessary for improving the WER.

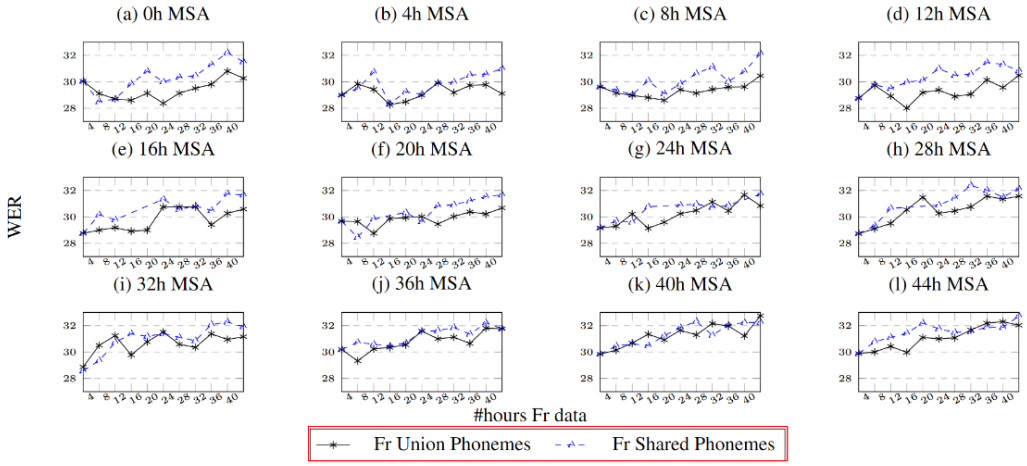

Figure 2. The WER variation on the ADIC Dev corpus for gradually extending ADIC Train corpus by MSA/French acoustic data.

5.3. Multitask learning

The principle idea of the multitask learning is to train one neural network with several sources of data to handle several tasks. For our purpose, we used the data of the three languages to train one model that recognises the three languages. Unlike the previous approach where all the layers of the neural network are shared among the three languages, in the multitask learning each language has a specific output layer, which means that the phonemes of each language are modelled separately. To update the parameters of the neural network, a simple way is to train it over different mini batches from each language. However, since we were not interested in the recognition of MSA nor French, the parameters of the neural network were updated by associating a weight w_l for each language in such a way that $\Sigma_{\{l=1\}}^{3} w_l = 1$ as applied in [21]. These weights were used to adjust the parameters of the hidden layers (λ_{sh}) after training over 400k samples of acoustic parameters according to the formula 1.

$$\lambda_{sh} = \sum_{l=1}^{3} w_l \lambda_{sh}^{l} \quad (1)$$

This is equivalent to train three models separately one for each language where each model has a set of parameters corresponding to the shared layers λ_{sh}^{l} and those of the output layer λ_{out}^{l}. The parameters of the shared layers in the global model λ_{sh} correspond to the weighted shared parameters of each model for each language.

To estimate the weights w_l for each language and in order to give more importance to the dialect ASR task, our training started with a high dialect weight (0.8) and it decreased gradually with a step of 0.1 in a such way that the dialect has always the high weight comparing to MSA and French. We opted for this approach because the training process is time consuming and it is hard to explore all the searching space. At the end of the estimation process, the weights that ensure better results on the Dev ADIC were found to be 0.5 for dialect and 0.5 for MSA if the model is trained on two tasks. If the system is trained on three tasks, the value of those weights was founded to be 0.4 for dialect, 0.3 for MSA and 0.3 for French.

5.4. Transfer learning

In the case where a small amount of data is available to train the neural network, it is common to pretrain a model on a large dataset and use it as a fixed feature extractor for the new task. In this case, hidden layers of the original network were fixed and a new task-specific layers were added over them. As in the multitask learning, phonemes between languages are not shared in this approach, but we need to find a way to update the model parameters. The common used approach is to update the parameters of the new added layers using a large learning rate (0.0005 in our case) and to fine-tune the parameters of original hidden layers with a small learning rate (0.00005 in our case).

To estimate the number of hidden layers n to transfer, an initial neural network was trained on MSA and French data by using our multilingual training approach. Afterwards, the output layer of this network is replaced by a specific layer for the dialect while keeping the n-first hidden layers. Those are $n \in \{1,2,3,4,5\}$ knowing that the initial model is composed of 6 hidden layers. The obtained WER on the Dev part of ADIC are presented in Figure 3. The results show that keeping the four first hidden layers from the initial model ensures better results.

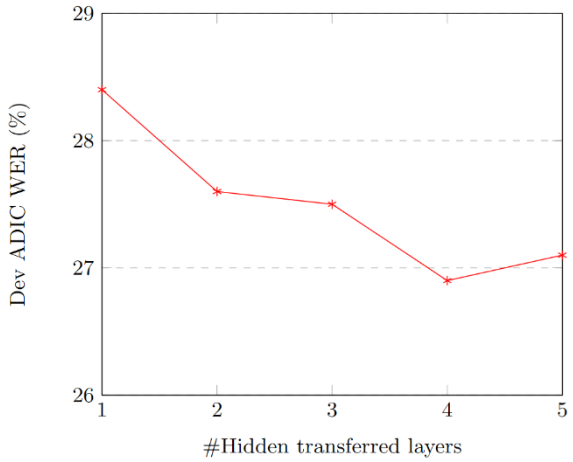

Figure 3. The impact of the number of hidden layers on the transfer learning.

6. RESULTS AND DISCUSSION

We used several corpora to train the acoustic model: MSA spoken data were extracted from NEMLAR[2] and NetDC[3] corpora, French data were extracted from ESTER corpus [22], and the dialect data are our recorded corpus ADIC. Forty dimensional MFCC feature vectors are used as input of the neural network at each timestamp. These MFCC features are extended with 100-dimensional identity vector (i-vector) [23]. I-vectors are low-dimensional vectors of speech segment used to describe the speaker characteristic in the speech. Despite that technique was initially proposed for speaker verification and speaker recognition tasks, it is also useful for speech recognition since it encapsulates the speaker relevant information in a low-dimensional representation. The implementation was based on Kaldi [24], a state-of-the-art toolkit for speech recognition based on weighted finite state transducers [25], and the experiments were carried out on Grid5000 platform [26].

6.1. Recognising the Algerian dialect using MSA-ASR system

We aim through this investigation to show how the Algerian dialect differs from the MSA. No dialectal data are used to train the language nor the acoustic models. The acoustic model is trained on 44 hours of MSA spoken data. We interpolate two bigram LMs trained on the MSA version of Gigaword corpus and on the transcripts of MSA speech training data; the interpolation weights are estimated on a MSA development corpus. The lexicon contains the most frequent words of the textual data used to train the LM. It has 95k unique words and 485k pronunciation variants. The results obtained with this ASR system are reported in Table 5. The test on MSA has been achieved on 5 hours of MSA speech data, while 1 hour and 15 minutes of the Test part of ADIC have been dedicated to test the ASR system on dialectal data.

Table 5. Performance of the MSA-ASR system on MSA and on the Algerian dialect.

System	Test	WER (%)	OOV (%)
MSA-ASR	MSA	12.7	2.5
	Test ADIC	76.3	33.6

[2] http://catalog.elra.info/product_info.php?products_id=874
[3] http://catalog.elra.info/product_info.php?products_id=13&language=fr

Whereas the MSA-ASR system performs well on MSA (a Word Error Rate (WER) of 12.7%), it collapses completely when it is tested on the dialectal corpus (a WER of 76.3%). The Out-Of-Vocabulary (OOV) rate shows how MSA and Algerian dialect are different. These results confirm that it is impossible to directly recognise the Algerian dialect with an ASR system developed for MSA.

We report in the Table 6 the recognition results when applied on the Test part of ADIC according to the proposed approaches.

Table 6. The Algerian dialect speech recognition results according to the way of using data from foreign languages.

Training Approaches	**Training data**			**WER (%)**	**OOV (%)**
	Acoustic	**Lexicon**	**Textual**		
Monolingual training	44hMSA	MSA	MSA	76.3	33.6
	4hDial	Dial	Dial	42.6	7.9
	4hDial	Dial+MSA	Dial+MSA	39.7	6.8
Multilingual training	4hDial+44hMSA	Dial+MSA	Dial+MSA	36.6	6.8
	Union 4hDial+44hMSA+44hFr			36.3	
	Shared 4hDial+44hMSA+44hFr			37.1	
	Union 4hDial+12hMSA+12hFr			35.9	
	Shared 4hDial+12hMSA+12hFr			38.5	
Multitask learning	4hDial+44hMSA (mini batch)	Dial+MSA	Dial+MSA	36.5	6.8
	4hDial+44hMSA (weights averaging)			37.0	
	4hDial+44hMSA+44hFr (mini batch)			36.6	
	4hDial+44hMSA+44hFr (weights averaging)			36.9	
Transfer learning	44hMSA (Initial model) => 4hDial	Dial+MSA	Dial+MSA	38.1	6.8
	44hMSA+44hFr (Initial model) => 4hDial			37.1	

6.2. Monolingual training

In this approach, we used data from only one language to train the dialectal acoustic model. We remark that the use of dialectal data to train the different model improves the WER of the MSA-ASR system by 33 points (76.3% vs. 42.6%). These results were expected because of two main reasons: firstly, we used data that was specific to our task and thus led to a low OOV rate. Secondly, the dialectal phonemes were well-modelled by the acoustic model. Better still, including MSA textual data in the language modelling improves the system by 2.9% (42.6% vs. 39.7%).

6.3. Multilingual training

The multilingual training approach aimed to take advantage from the speech data of other languages to improve the recognition of the Algerian dialect. The experiments started by integrating the MSA spoken data in the training process of the acoustic model. This leads to an

absolute improvement of 3.1% (39.7% vs. 36.6%) which shows how the MSA data are important in the acoustic and the language modelling of the dialect. However, integrating French data in the training process of the acoustic model of the Algerian dialect does not improve the WER. Even this poor improvement, we can remark that better results were obtained when the common phonemes between languages are modelled separately without any sharing. This could be explained by the fact that the shared phonemes between the MSA and the French languages, even if they are the same, they are used in different phonological contexts that makes their pronunciations different in each language. Consequently, they should be separated in order to ensure a good ASR system performance for the Algerian dialect. We also find that optimizing the size of the MSA and French acoustic data leads to a better result (a WER of 35.9%). This allow us to prevent the overfitting issue on MSA and French data.

6.4. Multitask learning

The neural network in the multitask learning was trained on several speech recognition tasks while allowing for the hidden layers to be shared and each task to have a specific output layer. We investigated two configurations according to the number of tasks to train. The first configuration aims to study the impact of MSA data on the speech recognition of the Algerian dialect by training the model on two tasks: ASR for MSA and for dialect. In the second configuration, the French ASR task was integrated in the training process.

For each configuration, we found that training the neural network over different mini batches from each language gives better results compared to the approach where we attributed weights for the different languages (weights averaging in Table 6). Knowing that the success of the weights averaging approach depends on the good estimation of the weights w_l for each language, we can explain the obtained results by our algorithm used to estimate these weights. In fact, we fixed a heigh weight for the dialect compared to the other languages; it would be interesting in this case to explore a large searching space where we fixed a low weight for the dialect.

The results also show that training the neural network on two tasks (ASR for MSA and for dialect) leads to an absolute improvement of 2.7% compared to the case where the neural network was trained on one task (ASR for dialect). This shows the importance of MSA data on the acoustic modelling of the Algerian dialect and confirms the obtained results in the multilingual training. However, we found that integrating the French ASR task in the training process brings no benefit to the system's performance.

6.5. Transfer learning

We aimed in the transfer learning to train initial models on MSA and/or French data, to retain the four first hidden layers and to add new dialect task specific layer over those.

We trained two neural networks using the multilingual training approach to study the impact of each language on the ASR of Algerian dialect. The first model was trained only on the MSA data while in the second one the French data were also integrated in the training process. These two models are used, afterwards, for the transfer learning.

Unlike what we found in the multilingual training and in the multitask learning, the French spoken data improve the system’s performance. The best WER was the one obtained by adapting the acoustic model trained on MSA and French data to the Algerian dialect.

By comparing our three approaches of integrating data from several languages into the training process of the acoustic model of the Algerian dialect, we found that the best approach is the one based on the multilingual training (a WER of 35.9%) where all layers of the neural network were shared between the three languages. This allows an implicit increase in the size of the data we used to train our model, which allows the model to better capture the relationship between the three languages and thus improve the dialect ASR system. It should be noted that the confidence interval for the system trained on the dialectal acoustic data only was ±1.65% (the one achieved

a WER of 39.7% in Table 6), which means that integrating MSA and French spoken data in the training process of the acoustic model for the dialect achieves a significant improvement.

There is relatively few research works on ASR for Algerian dialects in order to be able to compare our obtained results. However, in the last edition of the MGB challenge, MGB5 [27], there was a task about ASR for Moroccan dialect, which is relatively close to the Algerian dialect because they share several linguistic and acoustic aspects. The best system obtained a WER of 37.6%, knowing that 13 hours of dialectal speech were used with 1200 hours of MSA to train the acoustic model. This shows how hard the speech recognition task of Maghrebi dialects, especially the Algerian one, and that the results of our system are acceptable.

3. CONCLUSIONS

This work investigated developing an ASR system for Algerian dialect by starting from an ASR system dedicated to MSA. This attempt collapses completely when it was used to recognize the dialect (a WER of 76.3%). This shows how different are Algerian dialects and MSA.

To overcome the lack of spoken resources for this dialect and since Algerian dialects are mainly impacted by MSA and French languages, we investigated the use of acoustic data from these two languages. We showed that it is possible to develop an acoustic model on the base of a small recording dialectal corpus then adding it to larger corpora of well-resourced languages such as French and Arabic. It could be interesting to investigate this approach to develop ASR systems for other dialects especially those impacted by French such as Moroccan and Tunisian dialects. The recorded dialectal corpus provides a valuable resource for further studies on the Algerian dialect.

Through our investigation, we showed that sharing all layers of the neural network based acoustic model (the multilingual training) ensures best results compared to sharing only hidden layers (multitask and transfer learning). Furthermore, taking the union of phonemes of the three languages, in the case where the output layer is shared, led to a better acoustic model compared to the case of considering the intersection of common phonemes. We also investigated the required amount of data required to train a decent dialectal acoustic model. Our conclusion is that selecting subsets of data led to a better speech recognition system compared to using a larger amount of data. This is because with larger amounts of speech data from one of the mixture of languages, the performance can be impacted negatively. The over representation of a particular language makes the ASR system more sensitive to the phonemes of this language and less to the others.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of Chist-Era for funding part of this work through the AMIS (Access Multilingual Information opinionS) project.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

REFERENCES

[1] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard and A. De Waal, “A smartphone-based ASR data collection tool for under-resourced languages,” *Speech communication,* vol. 56, p. 119–131, 2014.

[2] F. de Wet, N. Kleynhans, D. Van Compernolle and R. Sahraeian, “Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems,” *South African Journal of Science,* vol. 113, p. 1–9, 2017.

[3] V.-B. Le and L. Besacier, “Automatic speech recognition for under-resourced languages: application to Vietnamese language,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, p. 1471–1482, 2009.

[4] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, L. Lamel, H. Maynard, M. Müller and others, “A very low resource language speech corpus for computational language documentation experiments,” *arXiv preprint arXiv:1710.03501*, 2017.

[5] D. Amazouz, M. Adda-Decker and L. Lamel, “Addressing Code-Switching in French/Algerian Arabic Speech,” in *Proceedings of Interspeech*, 2017.

[6] k. Abidi, M. a. Menacer and K. Smaili, “CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube,” in *18th Annual Conference of the International Communication Association (Interspeech)*, 2017.

[7] M. Killer, S. Stuker and T. Schultz, “Grapheme based speech recognition,” in *Eighth European Conference on Speech Communication and Technology*, 2003.

[8] H. Cucu, L. Besacier, C. Burileanu and A. Buzo, “Investigating the role of machine translated text in ASR domain adaptation: Unsupervised and semi-supervised methods,” in *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, 2011.

[9] P. Karanasou and L. Lamel, “Comparing SMT methods for automatic generation of pronunciation variants,” in *International Conference on Natural Language Processing*, 2010.

[10] S. Harrat, K. Meftouh, M. Abbas and K. Smaïli, “Grapheme to phoneme conversion - an Arabic dialect case,” in *Spoken Language Technologies for Under-resourced Languages*, 2014.

[11] A. Masmoudi, F. Bougares, M. Ellouze, Y. Estève and L. Belguith, “Automatic speech recognition system for Tunisian dialect,” *Language Resources and Evaluation*, vol. 52, p. 249–267, 01 3 2018.

[12] D. Yu, B. Varadarajan, L. Deng and A. Acero, “Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion,” *Computer Speech & Language*, vol. 24, p. 433–444, 2010.

[13] K. Kirchhoff and D. Vergyri, “Cross-dialectal acoustic data sharing for Arabic speech recognition,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[14] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas and K. Smaili, “Machine translation experiments on PADIC: A parallel Arabic dialect corpus,” in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, 2015.

[15] K. Meftouh, S. Harrat and K. Smaïli, “PADIC: extension and new experiments,” in *7th International Conference on Advanced Technologies ICAT*, Antalya, 2018.

[16] K. Abidi and K. Smaïli, “An automatic learning of an Algerian dialect lexicon by using multilingual word embeddings,” in *11th edition of the Language Resources and Evaluation Conference, LREC 2018*, 2018.

[17] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel and J. Glass, “A complete KALDI recipe for building Arabic speech recognition systems,” in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 2014.

[18] V. Peddinti, D. Povey and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, “Phoneme recognition using time-delay neural networks,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, p. 328–339, 1989.

[20] K. Veselỳ, A. Ghoshal, L. Burget and D. Povey, “Sequence-discriminative training of deep neural networks.,” 2013.

[21] R. Sahraeian and D. V. Compernolle, “Using Weighted Model Averaging in Distributed Multilingual DNNs to Improve Low Resource ASR,” *Procedia Computer Science*, vol. 81, pp. 152-158, 2016.

[22] S. Galliano, G. Gravier and L. Chaubard, “The ester 2 evaluation campaign for the rich transcription of french radio broadcasts,” in *Proceddings of Interspeech, Brighton (United Kingdom)*, 2009.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, p. 788–798, 2010.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, “The KALDI Speech Recognition Toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.

[25] M. Mohri, F. Pereira and M. Riley, “Speech recognition with weighted finite-state transducers,” in *Springer Handbook of Speech Processing*, Springer, 2008, p. 559–584.

[26] D. Balouek, A. Carpen Amarie, G. Charrier, F. Desprez, E. Jeannot, E. Jeanvoine, A. Lèbre, D. Margery, N. Niclausse, L. Nussbaum, O. Richard, C. Pérez, F. Quesnel, C. Rohr and L. Sarzyniec, “Adding Virtualization Capabilities to the Grid'5000 Testbed,” in *Cloud Computing and Services Science*, vol. 367, I. I. Ivanov, M. van Sinderen, F. Leymann and T. Shan, Eds., Springer International Publishing, 2013, pp. 3-20.

[27] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals and K. Choukri, “The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech,” 2019.