Automatic speech recognition for Tunisian dialect

Ahmed Ben Ltaief[1], Yannick Estève[2], Marwa Graja[1], and Lamia Hadrich Belguith[1]

ANLP Research group, MIRACL Lab., University of Sfax, Tunisia[1]
LIUM, Le Mans University, France[2]
ahmedbenltaief92@gmail.com
yannick.esteve@univ-lemans.fr
marwa.graja@gmail.com
l.belguith@fsegs.rnu.tn

Abstract. Speech recognition for under-resourced languages represents an active field of research during the past decade. The tunisian arabic dialect has been chosen as a typical example for an under-resourced Arabic dialect. We propose, in this paper, our first steps to build an automatic speech recognition system for Tunisian dialect. Several Acoustic Models have been trained using HMM-GMM and HMM-DNN system. The speech corpus has been collected and transcribed from dialogues in the Tunisian Railway Transport Network. The HMM-DNN system can give an impressive relative reduction in WER.

Keywords

Tunisian dialect, ASR system, HMM-GMM, HMM-DNN, Kaldi

1 Introduction

Automatic Speech Recognition (ASR) is playing an increasingly important role in a variety of applications. Nowadays, computers are heavily used to communicate via text and speech. While ASR is well developed for some langages such as english, it represents a challenging task with an under-resourced langage such as Arabic, due to lack of resouces (corpora, dictionary, ...).
Arabic is considered as one of the most morphologically complex languages. In fact, the Arabic dialect is a collection of spoken varieties of Arabic used in everyday life communications. It is only spoken and not formally written, and represents various degrees of differences in terms of phonology, morphology, syntax and lexicon.
In this paper, the tunisian arabic dialect has been chosen as a typical example for an under-resourced Arabic dialect. Tunisian arabic dialect is the primary dialect spoken in Tunisia and has unique features that distinguish it from the other Arabic dialects. However, Tunisian dialect still quite behind the state-of-the-art for MSA in NLP, even some other dialects. So the need for work on technologies for Tunisian dialect is more real than ever before.

The remainder of this paper is organized as follows: section 2 introduces Tunisian dialect. Section 3 presents the ASR and the different component of an ASR system. A

brief related works for MSA and dialect will be presented in section 4. In section 5, we will present our differents models for building an ASR system for Tunisian dialect and discussing the experimental results. Section 6 concludes this work.

2 Tunisian Dialect

Tunisian dialect is the spoken Arabic language in Tunisia. It is used in daily life and emerged as as the language of communication online: social media, blogs, SMS, emails, etc. It exists many regional varieties of Tunisian dialect and it differs from a region to another. This varieties includes the Tunis dialect (Capital), Sahil dialect, Sfax dialect, Northwestern Tunisian dialect, Southwestern Tunisian dialect, and Southeastern Tunisian dialect [3].
[13] classified the Tunisian dialect into four classes. The first one includes words derived from MSA roots with the application of MSA patterns. The second class gathers Words derived from Tunisian dialect roots via the application of the derivation patterns of MSA. The third class consists on applying TD specific patterns on words that are derived from the MSA roots. The last class includes words which are derived from foreign languages especially French and Italian.

3 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a tehnologie that converts an audio signal to text. Speech recognition software is now frequently installed in computers and mobile devices, allowing for easy access and makes life easier.
ASR has a wide range of applications like command recognition, dictation, interactive voice response, learning foreign language [10], automatic query answering, speech-to-text transcription, etc. It can be also beneficial for handicapped people to interact with society.
Researchers on automatic speech recognition have several potential choices of open-source toolkits for building a recognition system such as HTK [11], CMU Sphinx [6], Julius [5],Kaldi [9], etc. General architecture of an ASR system that uses the HMM-based approach is presented in figure 1.

The waveforme audio given in the input is converted into a sequence of fixed size acoustic vectors Y=y_1,...,y_n. The decoder tries then to find the sequence of words W=w_1,...,w_n which is which is most likely to have generated Y.

$\hat{W}$=arg$_w$max{P(Y|w)P(w)}

P(Y|w) is determined by an acoustic model and P(w) is determined by a language model.

3.1 Feature extraction

The first step in any ASR system is to extract features. It consists in identifying the component of an audio signal that are useful for the recognition task and discarding all

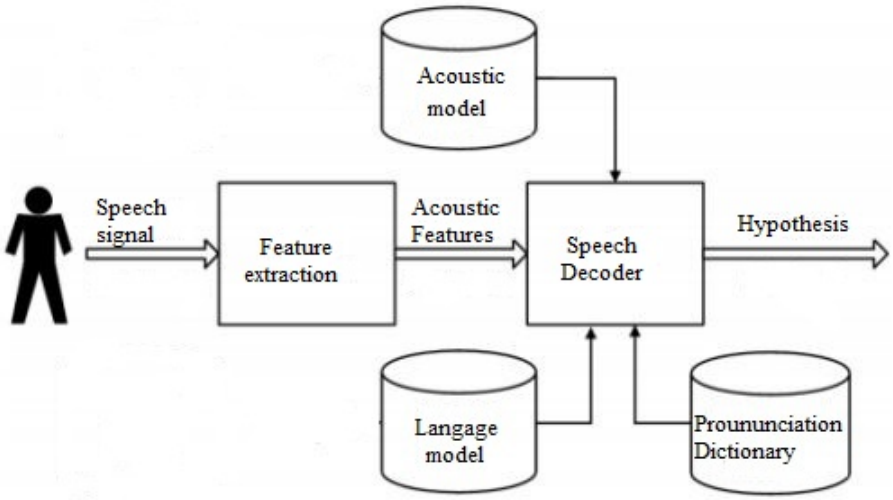

Fig. 1: ASR system architecture.

the remaining useless information such as noise, emotion, etc.
Mel Frequency Ceptral Coefficient (MFCC) is one of the common methods for feature extraction. The signal is framed into 20-40 ms frames. The intuition underlying this assumtion is that on short time scales the audio signal doesn't change much and we assume that the signal is enough stable. Afterwards, for each frame a vector of acoustic parameter is extracted.

3.2 Acoustic model

Acoustic modeling represents the relationship between linguistic units of speech and audio signals. An acoustic model is created by taking a large database of speech and using special training algorithms to create statistical representations for each phoneme in a language. These statistical representations are called Hidden Markov Models (HMM). Each phoneme has its own HMM.
A Hidden Markov Model (HMM) is a statistical model for representing probability distributions over sequences of observations. A HMM consists of two stochastic processes, an invisible process of hidden states and a visible process of observable events. HMMs models are trained on data with the forward-backward or Baum-Welch algorithm.
In speech recognition, the speech recognizer tries to find the sequence of phonemes (states) that gave rise to the actual uttered sound (observations).
The HMM relies on the assumption that the probability of being in a state at time t depends only on the state at time t-1. Formally:

$P(z_t|z_{t-1},z_{t-2}, ...)=P(z_t|z_{t-1})$

If the acoustic model is trained with sufficient number of speakers, it will be able to represent the properties of new speakers. This model is called speaker-independent. If the corpus includes the data of a specific speakers, such as a model trained with speakers belonging to the same community, it is considered speaker-dependent.

3.3 Pronunciation Dictionary

The Pronunciation Dictionary (PD) or lexicons has an important role in the predictive powers of ASR. It maps vocabulary words to sequence of phonemes which indicates the pronunciation of each of these words. For example: hello H EH L OW.
Lexicon should cover all the words we need, otherwise the system will not be able to recognize them. To fix this problem, we need a langage model (we will present it in the next section); the system looks for the word both in the lexicon and in the language model.

3.4 Langage Model

The Language Model (LM) is used to estimate the probability of a word given the word sequence that has already been observed:

$P(w_n|w_1,w_2,w_3...w_{n-1})$

We distinguish two types of langage model: Grammar Based Language Models (GBLM) and Probabilistic Language Models (PLM). The latter are currently dominate the field of speech recognition because GBLM is not generally useful for large vocabulary applications and it is so difficult to write a grammar with sufficient coverage.
N-gram models are the simplest kind of statistical language model. The basic idea is to consider the structure of a corpus as the probability of different words occurring alone or occurring in sequence.
3-gram are probably the most common ones used in ASR and represent a good balance between complexity and robust estimation. In a 3-gram language model the probability of a word given it's predecessors is estimated by the probability given the previous two words:

P(wn|w1,w2,w3,w4,...wn-1) = P(wn|wn-2,wn-1)

3.5 Speech decoder

Speech decoder is one of the central parts of ASR system. Decoding calculates which sequence of words is most likely to match best with speech given an acoustic and langage models. The decoder listens for the distinct sounds spoken by a user and then looks for a matching HMM in the Acoustic Model. If it is true, it keeps track of the matching

phonemes. Then, the decoder looks up the matching series of phonemes it finds in its Pronunciation Dictionary to determine which word is spoken.

4 Related works

[8] proposed an enhanced ASR system for Arabic (MSA). The author used Kaldi toolkit to build their system around. They trained different acoustic models using HMM-GMM system with several techniques such as LDA+MLLT, SAT, fMLLR, and HMM-DNN systems. Two corpora are used to train the acoustic model: Nemlar and NetDC consisting of 63 hours of Standard Arabic news broadcasts. The langage model was trained also with two corpora: GigaWord3 Arabic corpus and the acoustic training data transcription. The former has 1,000 million word occurrences and the latter has 315K words. the best result is obtained with DNN models achieving 14.42 of WER.

[1] described their ASR system for MSA using kaldi toolkit. They built three triphone models: a GMM-HMM, SGMM-HMM and DNN-HMM models. A preprocessing phase was integrated which does autocorrection of the original text represented in the normalization and the vowelization using MADA toolkit [4]. The dataset contains two types of speech: 127 hours of broadcast conversations and 76 hours of broadcast reports. The lexicon has 526K unique words, with 2M pronunciations, and The LM has 1.4M words.
The experiments showed better results when using a normalized text than a raw or normalized and vowelized text for the LM using the SGMM+bMMI AM. The best results are obtained using SGMM+bMMI,DNN and DNN+MPE having respectively 30.73%, 29.81% and 26.95%.

[2] proposed cross-lingual language approach for the development of a TV broadcasts system for Qatari dialect.The ASR system is a GMM-HMM architecture based on the speech recognition toolkit Kaldi [9].
Firstable, the Qatari LM was interpolated with the MSA-LM which is trained using the LDC Gigaword corpus. this interpolation resulted in a vocabulary size of 265.7K words. Thereby, a significantly decrease of out of vocablary (OOV) rate. Afterwards, a MSA acoustic model is used to decode qatari speech data. A data pooling technique was used after the previous step consisting on training an acoustic model using both qatari and MSA data.
Then, an acoustic aodel adaptation was applied using Maximum Likelihood Linear Regression (MLLR) and Maximum APosteriori (MAP) re-estimation on the MSA model using Qatari speech Data. A combination of different system was applied finally leading on 21.3% and 28.9% relative WER reduction on QA development set and evaluation set respectively. The MSA corpus consists of two speech resources: the NEMLAR Broadcast News Speech Corpus which consists of about 40 hours of audio, and the NetDC which has about 22.5 hours of Arabic Broadcast News Speech. The Qatari corpus consists of 15 hours collected from different TV series and talk show programs.

5 Experiments

We used Kaldi Recognition toolkit [9] to build our system. Kaldi has attracted speech researchers, and it has been very actively developed over the past few years. Kaldi is released under the Apache license v2.0, which is flexible and fairly open license. It includes recipes for training ASR systems with many speech corpora which are available and frequently updated with the latest techniques, such as Bottle-Neck Features (BNF), Deep Neural Networks (DNN), etc.

5.1 Dataset

We used for our experiments the TARIC corpus (Tunisian Arabic Railway Interaction Corpus) [7] which is a collection of audio recordings and transcriptions from dialogues in the Tunisian Railway Transport Network. The TARIC corpus was manually transcribed due to the absence of tools for automatic transcription for tunisian arabic. Then, a normalization step was applied to obtain coherent data using a standard orthographies described in [12].

Data	**#hours**	**#vocablary**
Train	8 Hours and 57 Minutes	3027
Dev	33 Minutes and 40 Seconds	612
Test	43 Minutes and 14 Seconds	1009

Table 1: Number of hours and vocablary for each dataset.

5.2 Acoustic Modeling

Different models are trained using HMM-GMM and HMM-DNN systems during acoustic model training. The first one was trained using MFCC, deltas and deltas-deltas features. Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) are applied to reduce dimensionality, which improves accuracy as well as recognition speed. Speaker Adaptive Training (SAT) was introduced afterwards with Feature-space MLLR (fMLLR).
A Maximum Mutual Information (MMI) criterion is used also to give a high discriminative ability to the system and thus MMI belongs to the so called "discriminative training" category.

Finally, we trained a DNN model on top of fMLLR features. The training is done in three stages: RBM pre-training, frame cross-entropy training which has as objective to classify frames to correct pdfs, and the sequence-training optimizing sMBR. The DNN training was done using one GPU on a single machine.

5.3 Langage Modeling

Our LM was built using TARIC transcription. We have trained a 3-gram model given the training corpus and the vocablary using SRILM toolkit. The given vocablary has 3211 unique words.

5.4 Results

Table 2 shows the obtained results from different models generated. GMM models with SAT_fMLLR gave a gain of almost 5% in test and 3% in Dev. The MMI training did not lead to an improvment of results compared to SAT+fmllr training, gave an increase of 1% WER in the dev and test set. As expected, the Deep Neural Network (DNN) models gave an impressive gain, with an WER of 25% and 36.8% in dev and test set respectively. The DNN models give us a nice gain of 6.5% in dev set compared to the first model, and 12% in test. Compared with the best GMM model (with SAT+fmllr), DNN gave an improvement of almost 4% and 7% respectively for dev and test.

Model	**Techniques**	**Dev**	**Test**
HMM-GMM	MFCC+deltas+deltas-deltas	31.5	48.8
HMM-GMM	LDA_MLLT	31.2	48.7
HMM-GMM	SAT+fMLLR	28.9	43.6
HMM-GMM	MMI	29.4	44.4
HMM-DNN	-	25	36.8

Table 2: WER of our models.

6 Conclusion

In this paper, we present our work on establishing Kaldi recipes to build Tunisian speech recognition system. Different Acoustic models have been trained using different techniques in order to increase system performances. The best results are coming from the training of DNN models, with an overall WER of 25% for dev set 36.8% for test. As a future work, we will ameliorate this results by increasing our data set, and using cross-lingual approach to take benifit from the other langages.

References

1. Ali, A.M., Zhang, Y., Cardinal, P., Dehak, N., Vogel, S., Glass, J.R.: A complete KALDI recipe for building arabic speech recognition systems. In: 2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014. pp. 525–529. IEEE (2014), https://doi.org/10.1109/SLT.2014.7078629

2. Elmahdy, M., Hasegawa-Johnson, M., Mustafawi, E.: Development of a tv broadcasts speech recognition system for qatari arabic. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
3. Gibson, M.: Dialect Contact in Tunisian Arabic: Sociolinguistic and Structural Aspects. University of Reading (1999), https://books.google.fr/books?id=iFO9GwAACAAJ
4. Habash, N., Rambow, O., Roth, R.: Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization (2009)
5. Lee, A., Kawahara, T., Shikano, K.: Julius – an open source realtime large vocabulary recognition engine. In: in EUROSPEECH. pp. 1691–1694 (2001)
6. Lee, K.F., Hon, H.W., Reddy, R.: An overview of the sphinx speech recognition system. IEEE Trans. Acoustics, Speech, and Signal Processing 38(1), 35–45 (1990), http://dblp.uni-trier.de/db/journals/tsp/tsp38.html#LeeHR90
7. Masmoudi, A., Khmekhem, M.E., Estève, Y., Belguith, L.H., Habash, N.: A corpus and phonetic dictionary for tunisian arabic speech recognition. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014. pp. 306–310. European Language Resources Association (ELRA) (2014), http://www.lrec-conf.org/proceedings/lrec2014/summaries/454.html
8. Menacer, M.A., Mella, O., Fohr, D., Jouvet, D., Langlois, D., Smaili, K.: An enhanced automatic speech recognition system for arabic. In: Proceedings of the Third Arabic Natural Language Processing Workshop. pp. 157–165. Association for Computational Linguistics, Valencia, Spain (April 2017), http://www.aclweb.org/anthology/W17-1319
9. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011), http://publications.idiap.ch/index.php/publications/showcite/Povey_Idiap-RR-04-2012, iEEE Catalog No.: CFP11SRW-USB
10. Satori, H., Harti, M., Chenfour, N.: Introduction to arabic speech recognition using cmusphinx system. CoRR abs/0704.2083 (2007), http://arxiv.org/abs/0704.2083
11. Young, S.J.: The HTK Hidden Markov model ToolKit: Design and philosophy. Entropic Cambridge Research Laboratory, Ltd 2, 2–44 (1994)
12. Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L.H., Habash, N.: A conventional orthography for tunisian arabic. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014. pp. 2355–2361 (2014), http://www.lrec-conf.org/proceedings/lrec2014/summaries/219.html
13. Zribi, I., Khemakhem, M.E., Belguith, L.H.: Morphological analysis of tunisian dialect. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013. pp. 992–996. Asian Federation of Natural Language Processing / ACL (2013), http://aclweb.org/anthology/I/I13/I13-1133.pdf