25

Arabic Dialectical Speech Recognition in Mobile Communication Services

Qiru Zhou[1] and Imed Zitouni[2]
[1]*Bell Labs, Alcatel-Lucent, NJ,*
[2]*IBM T.J. Watson Research Center*
USA

1. Introduction

We present in this chapter a practical approach in building Arabic automatic speech recognition (ASR) system for mobile telecommunication service applications. We also present a procedure in conducting acoustic modelling adaptation to better take into account the pronunciation variation across the Arabic speaking countries.

Modern Standard Arabic (MSA) is the common spoken and written language for all the Arab countries, ranging from Morocco in the west to Syria in the East, including Egypt, and Tunisia. However, the pronunciation varies significantly from one country to another to a degree that two persons from different countries may not be able understand each other. This is because Arabic speaking countries are characterized by a large number of dialects that differ to an extent that they are no longer mutually intelligible and could almost be described as different languages. Arabic dialects are often spoken rather than written varieties. MSA is common across the Arab countries, but it is often influenced by the dialect of the speaker. This particularity of the Arabic countries constitutes a practical problem in the development of a speech-based application in this region; suppose a speech application system is built for one country influenced by one dialect, what does it take to adapt the system to serve another country with a different dialect region? This is particularly challenging since resource to build accurate speaker independent Arabic ASR system for mobile telecommunication service applications are limited for most of the Arabic dialects and countries.

Recent advances in speaker independent automatic speech recognition (SI-ASR) have demonstrated that highly accurate recognition can be achieved, if enough training data is available. However, the amount of available speech data that take into account the dialectal variation of each Arabic country is limited, making it challenging to build a high performance SI-ASR system, especially when we target specific applications. Another big challenge when building an SI-ASR is to handle speaker variations in spoken language. These variations can be due to age, gender, educational level as well as the dialectical variants of Arabic language. Usually an ASR system trained in one regional variation exhibits poorer performance when applied to another regional variation. Three problems may arise when a SI-ASR system built for one dialect but applied to target users with a different dialect: (1) Acoustic model mismatch, (2) Pronunciation lexicon mismatch and (3) Language model mismatch.

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

In the following we show how to build an Arabic speech recognition system for mobile telecommunication service applications. Furthermore we show how to adapt the acoustic model and the pronunciation lexicon of ASR systems for the Modern Standard Arabic to better take into account the pronunciation variation across the Arabic countries. This is a challenging problem especially when not enough data is available for every country. One of the topics we address is the dialect adaptation across the region. We investigate how a Modern Standard Arabic ASR system trained on one variation of the language in one country A can be adapted to perform better in a different Arab speaking country B, especially where only small amount of data related to country B is available. We show in this chapter how we take into account the pronunciation variation and how we adapt the acoustic model.

Experiments are conducted on Orientel (Oriental, 2001) database covering, in addition to Modern Standard Arabic, Arabic dialect spoken in Morocco, Tunisia, Egypt, Jordan, and United Arab Emirate. Results show an interesting improvement is achieved by using our adaptation technique. In this work, we experiment dialect adaptation from Tunisian (Maghreb Arabic) dialect to Jordan (Levantine Arabic) MSA dialect.

In section 2 and section 3, we discuss SI-ASR model training and adaptation techniques that are language neural. In section 4, we demonstrate a real-world practise on building Arabic speech recognition systems with both model estimation techniques and adaptation techniques. We then conclude the paper in section 5.

2. Hidden Markov Model parameter estimation for speaker independent speech recognition

Although other pattern recognition methods have been developed in the history of speech recognition research and development, the hidden Markov model (HMM) method is by far the most successful method used in speech recognition. Almost all modern speech recognition research and commercial systems are using some form of HMM to model the spectral and temporal variations of basic speech units. HMM is a very powerful statistical method of characterizing the observed data samples of a discrete-time series. For speaker independent speech recognition, HMM provides an efficient way to build parametric models on large amount of observation samples (e.g., speech data collection from many speakers.). Incorporate with the dynamic programming principle, it can be used for pattern segmentation and classification of a time-varying sequence. HMM and dynamic programming are the two key technologies for most of the modern SI-ASR systems today.

2.1 Hidden Markov Model

The basic HMM theory for pattern classification was developed by Baum and his colleagues during 1960s and 70s (Baum 1966~1970). The hidden Markov model is a statistical model that uses a finite number of states where the output observation is a random variable X generated according to a output probabilistic function associated with each state. It can be viewed as a double-embedded stochastic process with an underlying stochastic process (the state sequence) not directly observable (hence "hidden"). A hidden Markov model is defined as:

1. $\boldsymbol{O} = \{o_1, o_2, \cdots, o_M\}$ −An output observation alphabet.
2. $\boldsymbol{S} = \{s_1, s_2, \cdots, s_N\}$ − A set of states representing the state space of the model.

3. $A = \{a_{ij}\}$ −A transition probability matrix, where a_{ij} is the probability of taking a transition from state i to state j:

$$a_{ij} = P(s_t = j \mid s_{t-1} = i) \tag{1}$$

4. $\boldsymbol{B} = \{b_i(k)\}$ −An output probability matrix, where $b_i(k)$ is the probability of emitting symbol o_k when state I is entered. Let $\boldsymbol{X} = X_1, X_2, \cdots, X_t, \cdots$ be the observed output of the HMM. The state sequence $S = s_1, s_2, \cdots, s_t, \cdots$ is not observed (hidden). Therefore, $b_i(k)$ can written as:

$$b_i(k) = P(X_t = o_k \mid s_t = i) \tag{2}$$

5. $\boldsymbol{\pi} = \{\pi_i\}$ −An output initial state distribution, where

$$\pi_i = P(s_0 = i), \qquad 1 \le j \le N \tag{3}$$

All probabilities must satisfy the following properties

$$a_{ij} \ge 0,\ b_i(k) \ge 0, \pi_i \ge 0,\ \forall \text{ all } i, j, k \tag{4}$$

$$\sum_{k=1}^{N} a_{jk} = 1,\ \sum_{k=1}^{M} b_j(k) = 1,\ \sum_{i=1}^{N} \pi_i = 1 \tag{5}$$

In continuous observation density HMMs, the observations are continuous signals (vectors), the observation probabilities then often be replaced by finite mixture probability density functions (pdf):

$$b_j(o) = \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, \sigma_{jk}), \qquad 1 \le j \le N \tag{6}$$

where o is the observation vector being modelled. c_{jk} is the mixture coefficient for the kth mixture in state j and N is any log-concave or elliptically symmetric density (in speech recognition, Gaussian pdf is commonly used). Without loss of generality, we can assume that N is Gaussian in (3) with mean vector $\boldsymbol{\mu_{jk}}$ and covariance matrix $\boldsymbol{\sigma_{jk}}$ for the kth mixture component in state j. The mixture gain c_{jk} satisfy

$$\sum_{k=1}^{M} c_{jk} = 1, 1 \le j \le N \tag{7}$$

$$c_{jk} \ge 0, 1 \le j \le N, 1 \le k \le M \tag{8}$$

so that the the pdf is properly normalized, i.e.,

$$\int_{\infty}^{\infty} b_j(o) do = 1, \qquad 1 \le j \le N \tag{9}$$

In practical acoustic speech signal modelling, we generally assume that HMM is in left to right format with equal transitional probability

$$a_{ij} = a_{ii} = 1/2, \quad 1 \le i < N, i < j \tag{10}$$

$$a_{ij} = 0, \quad j > i + \Delta i \tag{11}$$

$$a_{NN} = 1, a_{Ni} = 0, \ i < N \tag{12}$$

and π is uniform distribution to simply the model topology. Hence the model parameter estimation becomes estimate probability matrix in equation (2) for discrete HMMs, and Gaussian pdf parameters in equation (6) for continuous density HMMs, given training speech data set and model topology. A typical 3-state left to right HMM phoneme model topology is shown in Figure 1.

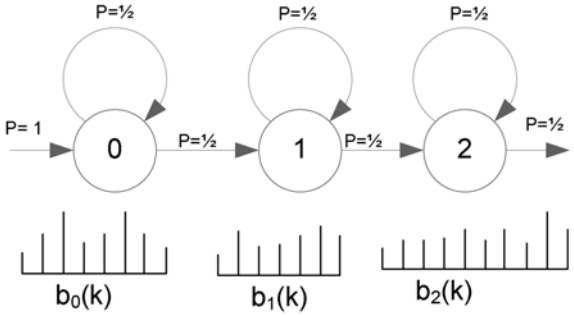

Fig. 1. A 3-state hidden Markov phoneme model topology example

2.2 HMM parameter estimation

Given the above form HMM, there are three basic problems of interests must be solved to for the HMM to be useful in real-world applications (Rabiner, 1989)

1. **The evaluation problem**
   Given the observation sequence $\boldsymbol{O} = (o_1, o_2, \ldots, o_T)$, and a model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, how do we compute $P(\boldsymbol{O} \mid \lambda)$, the probability of the observation sequence?
2. The decoding problem
   Given the observation sequence $\boldsymbol{O} = (o_1, o_2, \ldots, o_T)$ and the model λ, what is the most likely state sequence $\boldsymbol{q} = \{q_1, q_2 \ldots q_T\}$ in the model that produces the observations?
3. The learning problem
   How do we adjust the model parameters $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ to maximize $P(\boldsymbol{O} \mid \lambda)$?

In this chapter, we concentrate on the third problem which is the problem on how to train HMMs, given a set of speech training data and model topology. Combine with the forward and backward procedure designate to solve Problem 1 with Baum-Welsh EM (expectation-maximization) method (Dempster, 1977) using maximum likelihood (ML) approach, an iterative procedure is developed to estimate and re-estimate continuous density HMM model parameters efficiently (Rabiner, 1989):

$$\overline{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{13}$$

$$\overline{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \bullet O_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{14}$$

$$\overline{\sigma}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \bullet (O_t - \mu_{jk})(O_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{15}$$

where $\gamma_t(j, k)$ is the probability of being in state j at time t with the k^{th} mixture component account for o_t:

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \right] \left[\frac{c_{jk} N(o, \mu_{jk}, \sigma_{jk})}{\sum_{m=1}^{M} c_{jm} N(o, \mu_{jk}, \sigma_{jk})} \right] \tag{16}$$

Forward variable

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = i \mid \lambda) \tag{17}$$

Can be calculated efficiently using inductive algorithm:

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \le \mathrm{i} \le \mathrm{N} \tag{18}$$

2. Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix} \tag{19}$$

3. Termination

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{20}$$

Similarly, the backward variable

$$\beta_t(i) = P(o_{t+1} o_{t+2} ... o_T \mid q_t = i, \lambda) \tag{21}$$

Can be calculated:

1. Initialization

$$\beta_T(i) = 1, 1 \le \mathrm{i} \le \mathrm{N} \tag{22}$$

2. Induction

$$\beta_t(j) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \text{ , t = T-1, T-2, …,1, } 1 \le i \le N \tag{23}$$

Another commonly used approach is segmental ML training (also referred as Viterbi training)(Rabiner, 1986) which using K-means model parameter estimation and Viterbi algorithm re-segment the state sequence iteratively until model convergence.
In following sections, we discuss issues to train real-world application continuous density HMMs. Similar procedure applies to HMM training with tied mixture density functions and tied-state HMMs, with moderate algorithm modifications.

2.3 HMM topology

In HMM speech modelling, we assume each of the HMM state capture part of the speech as a quasi-stationary segment (20 ~ 30 milliseconds). For left-to-right HMM, all we need to decide is the number of states. This decision is depend on the word pronunciation and duration of the speech the HMM to model. For a long word or a phrase model, a HMM to model it may need 10 or more states. For a sub word or a phoneme model, 3 to 5 states are commonly used. Since silence and pause in speech is stationary, 1 to 3 states will be sufficient. Also for training purpose, available speed data to train the model will affect the number of states decision since there must be sufficient data (speech feature vectors) to train each of the states.

2.4 Initial estimation

A key question is that how to choose initial estimates of the HMM parameters so that the local maximum is equal to or as close as possible to the global maximum of the likelihood function? Experimentally, we observed that for raw training data, random or uniform initial estimates of $\boldsymbol{A}$ (Equation (1)) and $\boldsymbol{\pi}$ (Equation (3)) are adequate but a good initial estimation of $\boldsymbol{B}$ is required for continuous density HMM estimation. Here the most difficult problem is model state level segmentation that is hard to do by human labelling process. Here are a few practices in real-world application cases:

1. If there is a pre-built HMM that closes to the language acoustics of the target HMM, the pre-built HMM can be used to perform initial segmentation of the data to get better word, phoneme, and state boundaries. In case of acoustic feature mismatch, a signal and model conversion and rescaling is required.
   In this chapter, we will show how to use HMM trained on studio recorded, wideband speech to bootstrap HMM to be trained using narrow band telephony and mobile speech.
2. If there is human labelled speech data at speech element level (phoneme for sub word system, word for whole word recognition system.), we can use a random or uniform segmentation approach estimated the state segmentation to initialize the training. Then K-mean clustering ((MacQueen, 1967), (Hartigan & Wong, 1979)) can be used to create initial model. We can further refine the model using segmental kmeans iteration method (Rabiner 1986) or EM algorithms described above.
3. Furthermore, model adaptation and cross dialect and/or language phoneme mapping may be used for bootstrapping as well.

2.5 Effects of insufficient training data

In general, there are always finite amount of data to train HMM for a given target language and/or dialect. Hence we find always an inadequate number of occurrences of low probability events to give good estimates of the model parameters. There are literature discusses on effects of inadequate data and solutions on this issue ((Jelinek, 1980), (Rabiner & Juang 1992)). Here are some practices to address this issue:

1. Try to increase the training data to fill the data gap. But this is not always possible due to data collection availability or associated cost.
2. Decrease the model complexity (reduce model set size to remove rare seen acoustic event models, or reduce the mixture number for HMM with low count of available acoustic event). But this may produce model set with inadequate coverage of a language.
3. Using a floor value to replace the observation probability or density function.
4. For context dependent HMM, using tied state and furthermore tied mixture techniques to cover certain acoustic context events with low count.

Other smoothing techniques such as deleted interpolation (Jelinek, 1980) and Bayesian smoothing have been suggested to compensate the sparse data difficulties.

2.6 Duration modeling

Although HMM is a powerful statistical modelling tool, there are limitations of HMM to model real-world speech signals. One of the well known limitations is the duration model, which is the exponential decrease as represented as

$$d_i(t) = a_{ii}^t(1 - a_{ii}) \tag{24}$$

To improve duration modelling, an explicit time duration distribution can be built for each state. This duration distribution function parameters can be estimated from training data. Or a simple histogram distribution can be created which limit to finite number of duration time length.

3. Adaptation and corrective training yechniques

In real-world speech applications, it often requires a speech recognition system to adapt its acoustic and language models to new situations. Most of the speaker dependent system do need adapt to it's user in order to be produce satisfactory performance. Even a well trained robust speaker independent system that can accommodate wide range of speakers and environment may always have the situation that there is a mismatch between the model and the operating condition. If there is adequate training data and computing resource, the SI-ASR system may be re-trained to the new environment and users. But in most of the real-world situations, there is neither enough data nor resource to retrain the system. An alternative solution is to apply adaptation algorithms using limited data and computing resource to reduce the mismatches mentioned above to improve system performance. Therefore, a well built real-world speech recognition should have adaptation capability to minimize the possible mismatch in short time and minimum calibration data (e.g., a few utterances from a speaker can adapt the system in favour of the speaker using it.).

Many effective adaptation techniques have been developed to improve real-world speech applications. Basically, these adaptation techniques can be divided to two categories: model

adaptations and channel adaptations. The former adaptation changes acoustic and/or language model parameters (linear or non-linear transformations) to improve recognition accuracy. It is more suitable for speaker variations, unseen language situations and accents not covered in model training data. In most of the case, the system apply model adaptation are tuned to a specific speaker or a new dialect group of speakers. The latter is mainly addressing acoustic channel environment situation changes and improves recognition by tuning the system to be more environments robust. Channel adaptation (or front end adaptation) algorithms such as dynamic cepstral mean subtraction and signal bias removal (Rahim, 1994) are become a standard component to most of the modern speech recognition systems.

In this chapter, we are focusing on the former adaptation to address situation changes require model parameter changes for new dialects and vast acoustic environment changes. The most common adaptation techniques are maximum a posterior (MAP) and maximum likelihood linear regression (MLLR) algorithms. We briefly describe these two adaptation methods in the following sections.

3.1 MAP model adaptation

Maximum a posterior (MAP) estimation uses Bayesian learning framework to obtain estimation of random HMM parameter vector λ (Lee,1996). For a given set training / adaptation data $\mathbf{x}$, the conventional ML estimation assumes that λ is fixed but unknown and solves the equation

$$\lambda_{ML} = \arg\max_{\lambda} f(x \mid \lambda) \tag{25}$$

where f ($\mathbf{x} \mid \lambda$) is the likelihood of observation $\mathbf{x}$. i.e., MAP formulation assumes that the parameter λ to be a random vector with a know distribution f. Furthermore, we assume there is a correlation between the observation vectors and the parameters so that a statistical inference of λ can be made using a small set of adaptation data $\mathbf{x}$. Before making any new observations, λ is assumed to have a prior density g(λ) and new data are incorporated, λ is characterized by a posterior density g($\lambda \mid$x). The MAP estimate maximizes the posterior density

$$\lambda_{MAP} = \arg\max_{\lambda} g(\lambda \mid x) = \arg\max_{\lambda} f(x \mid \lambda) g(\lambda) \tag{26}$$

Since the parameters of a prior density can also be estimated from an existing HMM λ_0, this framework provides a way to combine with newly acquired data $\mathbf{x}$ in an optimal way.

Let $\mathbf{x} = (x_1, \ldots, x_N)$ be a set of scalar observations that are independent and identical distributed (i.i.d.) Gaussian distribution with mean m and variance σ^2. Here assume that the mean m is a random variable and the variance σ^2 is fixed. I can be shown that the conjugate prior for m is also Gaussian with mean μ and variance κ^2. If we use the conjugate prior for the mean to perform MAP adaptation, then the MAP estimation for the parameter $\mathbf{m}$ is

$$\tilde{m} = \frac{N\kappa^2}{\sigma^2 + N\kappa^2}\bar{x} + \frac{\sigma^2}{\sigma^2 + N\kappa^2}\mu \tag{27}$$

where N is the total number of training samples and $\bar{x}$ is the sample mean.

Using a prior density

$$g(\sigma^2) = \begin{cases} cons\tan t, if\sigma^2 \geq \sigma^2_{\min} \\ 0, otherwise \end{cases} \tag{28}$$

The MAP estimate of the variance is

$$\tilde{\sigma}^2 = \begin{cases} S_x, ifS_x \geq \sigma^2_{\min} \\ \sigma^2_{\min}, otherwise \end{cases} \tag{29}$$

where S_x is the sample variance of **x**.
From (27), we can see that the MAP estimation of the Gaussian mean is a weighted average of the prior mean μ and the sample mean.
The MAP training can be iterative as well. This requires an initial estimate of model parameters.

3.2 MMLR model adaptation

We can use a set of linear regression transformation functions to map both mean and covariance (variance for diagonal covariance matrix) in order to maximize the likelihood of the adaptation data (Leggetter, 1995). Since the transformation parameters can be estimated from relatively small amount of adaptation data, it is very effective for rapid adaptation. The maximum likelihood linear regression (MLLR) has been widely used to obtain adapt models for either a new speaker or a new environment condition.
Specifically, MLLR is a model adaptation method that estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. The transformation matrix used to give a new estimate of the adapted mean is given by

$$\tilde{\mu} = W\xi \tag{30}$$

Where W is the n x (n+1) transformation matrix (n is the dimensionality of the data) and ξ is the extended mean vector

$$\xi\hat{} = (\mathrm{w}\ \mu_1 \mu_2 \ldots \mu_\mathrm{n})^\mathrm{T} \tag{31}$$

where w is a bias offset (normally a constant). It has been show that W can be estimated as

$$w_i = k^{(i)} G^{(i)-1} \tag{32}$$

where w_i is the i^{th} row of W. and

$$G^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma^2_{mi}} \xi_m \xi_m^T \sum_{t=1}^{T} L_m(t) \tag{33}$$

and

$$k^{(i)} = \sum_{m=1}^{M} \sum_{t=1}^{T} L_m(t) \frac{1}{\sigma^2_{mi}} o_i(t) \xi_m^T \tag{34}$$

where $L_m(t)$ is the occupancy probability for the mixture component m at time t.
Similarly, variance transformation matrix can be calculated iteratively.

3.3 Comparison of MAP and MLLR

In general, MLLR performs better when adaptation data is limited. But when more adaptation data are available, MAP becomes more accurate. Also MLLR can combine with MAP to improve performance for both less and more adaptation data situations. When more adaptation data is available, MAP is more computation efficient. For more details, readers may refer to (Wang et al., 2003).

4. Experiments

In this section, we present and discuss our experiment on building SI-ASR for Arabic speech recognition for dialect variations. In this work, we use phoneme level sub word HMM models for the SI-ASR. Therefore the generic techniques on model estimation and adaptation all can apply to this work. The main purposes in this work is to bootstrap SIASR systems for telecommunication network applications, from a SI-ASR system built on studio recorded speech. Obviously, there is a significant acoustic mismatch and dialect mismatch. Another problem we need to solve is the phonetic mapping of the different systems. First, we introduce the speech corpora used in the experiment.

4.1 Speech corpora in experiments

The following speech copora are used in our experiments

4.1.1 West Point speech corpus

Since we don't have an initial bootstrap wireless telephony acoustic model, we bootstrap the experiment using an existing Arabic acoustic model trained on West Point Speech Corpus (WPA) (LaRocca, 2002) which is a collection of wideband studio speech recording.
The West Point Arabic Speech Corpus (WPA) contains MSA speech data that was collected by members of the Department of Foreign languages at the United States Military Academy at West Point and the Center for Technology Enhanced Language Learning (CTELL). The original purpose of this corpus was to train acoustic models for automatic speech recognition that could be used as an aid in teaching Arabic language. The corpus consists of 8,516 speech files, total 11.42 hours of speech data. Each speech file represents one person reciting one prompt from one of four prompt scripts. The utterances were recorded using a Shure SM10A microphone and a RANE Model MS1 pre-amplifier. The files were recorded as 16-bit PCM audio files, with a sampling rate of 22.05 KHz. Approximately 7200 of the recordings are from native informants and 1200 files are from non-native informants. Overall, there are about 30% of non-native speakers contributed about 10% of the total speech recording. There is no information given on the origin countries/regions of the native speakers in the database. There is 1128 distinct orthographic word in the WPA lexicon. The WPA phoneme symbol set is shown in Table 1.

4.1.2 Orientel corpora

OrienTel (OrienTel, 2001) is a data collection and research project driven by an international industrial and academic consortium. The goal of OrienTel is to enable the project's

Symbol	Description	Symbol	Description
C	voiced pharyngeal fricative	iy	high front tense vowel
D	velarized voiced alveolar stop	j	voiced palato-alveolar fricative
G	voiced velar fricative	k	voiceless velar stop
H	voiceless pharyngeal fricative	l	voiced alveolar lateral
Q	voiceless glottal stop	m	voiced bilabial nasal
S	velarized voiceless alveolar fricative	n	voiced alveolar nasal
T	velarized voiceless alveolar stop	q	voiceless uvular stop
TH	velarized voiced interdental fricative	r	voiced alveolar flap
Z	voiced interdental fricative	s	voiceless alveolar fricative
ae	low front vowel	sh	voiceless palato-alveolar fricative
ah	low back vowel	sil	silence
aw	back upgliding diphthong	sp	short pause
ay	front upgliding diphthong	t	voiceless alveolar stop
b	bilabial voiced stop	th	voiceless interdental fricative
d	voiced alveolar stop	uw	high back rounded vowel
ey	upper mid front vowel	w	voiced bilabial approximant
f	voiceless labiodental fricative	x	voiceless velar fricative
g	voiced velar stop	y	voiced palatal approximant
h	voiceless glottal fricative	z	voiced alveolar fricative
ih	high front lax vowel		

Table 1. West Point Arabic Speech Database Phoneme Set

participants to design and develop speech-based multilingual interactive communication services for the Mediterranean and the Middle East, ranging from Morocco in the West to the Gulf States in the east, including Turkey and Cyprus. These applications will typically be implemented on mobile and multi-modal platforms such as GSM or UMTS phones, personal digital assistants (PDAs) or combinations of the two.

Examples of applications are unified messaging, information retrieval, customer care, banking, and service portals. To achieve this goal, the consortium conducts various surveys of the OrienTel region, compiles a set of linguistic databases, conducts research into ASR-related problems the OrienTel languages hold and develops demonstrator applications bearing evidence of OrienTel's multilingual orientation.

During OrienTel project, 21 databases (Zitouni et. al., 2002) were collected to cover the four broad Arabic dialect regions, namely Mahgreb Arabic, Egyptian Arabic, Levantine Arabic and Gulf Arabic. Additional languages of commercial interest were also collected in the OrienTel region are English, French, German, Cypriote Greek, Turkish and Hebrew. Three

Country	1st linguistic variety	2nd linguistic variety	3rd linguistic variety	Patner
United Arab Emirates	Colloquial Gulf Arabic as spoken in the UAE	Standard Arabic as spoken in the UAE	English	Scansoft (Now Nuance)
Jordan	Colloquial Levantine Arabic as spoken in Jordan	Standard Arabic as spoken in Jordan	English	Lucent Technologies (Now Alcatel-Lucent)
Egypt	Colloquial Egyptian Arabic	Standard Arabic as spoken in Egypt	English	IBM
Morocco	Colloquial Maghreb Arabic spoken in Morocco	Standard Arabic as spoken in Morocco	French	ELDA/UPC
Tunisia	Colloquial Maghreb Arabic spoken in Tunisia	Standard Arabic as spoken in Tunisia	French	UPC/ELDA
Israel, Palestine	Colloquial Levantine Arabic as spoken in Israel and Palestine	Hebrew as spoken in Israel		NSC
Cyprus	Greek as spoken in Cyprus	English as spoken in Cyprus		Knowledge /Univ. of Patras
Turkey, Germany	Turkish as spoken in Turkey	German spoken by Turks in Germany		Siemens

Table 2. Collected linguistic varieties in the OrienTel region and patners

languages were collected for each of the Arabic countries: 500 speaker collection of the Modern Standard Arabic (MSA), 1000 speakers' collection of the Modern Colloquial Arabic (MCA), and 500 speakers of the 3rd language used in business in this country (English, French, German, etc.). The databases produced are shown in Table 2. The general speech contents classes of the databases are defined in Table 3. All OrienTel collections are phonetically labelled with SAMPA (SAMPA, 2005) phoneme system and symbol set (c.f. Table 5).

Two of the Orientel speech databases used in our study are MSA Tunisia and MSA Jordan (c.f. Table 4). The common features for the data collections are as the following: Each of the speakers recorded 51 utterances in his/her session, within the contents defined in Table 1. The total speech duration approximates to 45 hours for each of the databases. The recordings were performed from offices, homes, public places and on streets. The speech data is sufficient for acoustic training and testing for various application domains. About 80% of the data is defined as the training set and 20% is defined as the testing set. The division of the training set and testing set is based on recording sessions. Therefore no

speaker is in both sets. Some of the content classes suitable for training and other linguistic research (the A, W, S, X contents in Table 3) are excluded from testing set. Also training and testing sets were divided as even as possible per gender and age group distributions. The speaker ages are in the range of 16 to 60. The lexicon size for both of the databases is around 26,000.

Code	# of Uttr.	Utterance Description
I, B	3	isolated digits
C	8	digit/number strings
N	2	natural number
M	2	currency amount (local and foreign)
Q	2	yes/no questions
D	3	Dates
T	1	Times
A	6	application keyword/keyphrases
E	1	word spotting phrases
O	3	directory assistance names
L	2	Spellings
W	4	phonetically rich words
S	9	phonetically rich sentences
X	5	spontaneous items (for control use)

Table 3. OrienTel Database Content Definitions

MSA database	Speakers (Males/Females)	Dialect Region	Mobile Network (%)
Tunisia	598 (359/239)	Maghreb	70.0
Jordan	556 (288/268)	Levantine	70.5

Table 4. Orientel Tunisia and Jordan MSA Databases

All OrienTel speech corpora are collected mainly on mobile telephone networks with smaller portion of speech collected on wired line telephone networks. All the telephone networks are digital.

4.2 Experimental results

Our experiment has the following steps:

1. Speech data feature extraction and acoustic model structure selection.
2. Bootstrap using existing, wide band MSA ASR system to a narrow band ASR system.
3. Retrain and global MLLR adaptation to Tunisian MSA ASR system.
4. Dialect adaptation to Jordan MSA ASR system.

We illustrate in Figure 3 the different steps we follow in the next few sub-sections in order to build an accurate SI-ASR system for Jordanian MSA. We remind the reader that our goal is to show how we can adapt an Arabic MSA SI-ASR system built on speech data spoken in one country (e.g., Tunisia) to be more effective when used by speakers in another country (e.g., Jordan). Here we assume that we don't have enough data in the target language (e.g., Jordan) to train a complete system from scratch. This is why we consider the Tunisian MSA

SAMPA Symbol	WPA Symbol	SAMPA Keyword	Arabic Orthography	SAMPA Symbol	WPA Symbol	SAMPA Keyword	Arabic Orthography
Consonants: Plosives				**Nasals**			
b	b	ba:b	باب	m	m	ma:l	مال
t	t	tis?`	تسع	n	n	nu:r	نور
d	d	da:r	دار	**Trill**			
t`	T	t`a:bi?`	طابع	r	r	rima:l	رمال
d`	D	d`arab	ضرب	**Lateral**			
k	k	kabi:r	كبير	l	l	la:	لا
g	g	gami:l	جميل	l`	l	?al`l`ah	الله
?	Q	?akl	أكل	**Semivowels**			
p	-	paris	برس	w	w	wa:hid	واحد
Consonants: Fricatives				j	y	jawm	يوم
f	f	fi:l	فيل	**Vowels**			
v	-	nivi:n	نفين	i	ih	D`il	ظل
T	th	Tala:T	ثلاث	a	ah	X\al	حل
D	TH	Dakar	ذكر	u	uw	?`umr	عمر
D`	Z	D`ala:m	ظلام	i:	iy	?`i:d	عيد
s	s	sa?`i:d	سعيد	a:	ae	ma:l	مال
z	z	zami:l	زميل	u:	uw	fu:l	فول
s`	S	s`aGi:r	صغير	-	aw		
S	sh	Sams	شمس	-	ay		
Z	j	Zami:l	جميل	-	ey		
x	x	xit`a:b	خطاب				
G	G	Garb	غرب				
X\	H	X\ilm	حلم				
?` (?\)	C	?`alam	علم				
h	h	hawa:?	هواء				

Table 5. SAMPA to WPA phone set mapping

SI-ASR system as our baseline. The third step described earlier in this section shows how to built the Tunisian MSA SI-ASR system and the fourth step describes how the adaptation is conducted to build a more accurate SI-ASR system used by speakers from the target language (e.g. Jordanian).

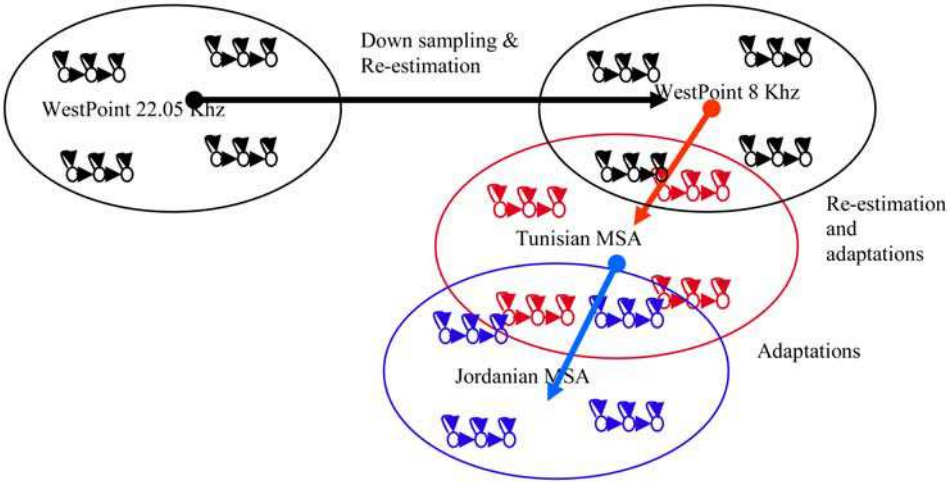

Fig. 3. Arabic Acoustic Model Dialect Adaptation

4.2.1 Feature extraction and model structure selection

Our experiment speech data feature extraction and model structure selections are as the following:

- We select 39 MFCC feature vector for speech feature extraction (12 c, 12 Δc, 12 ΔΔc, e, Δe, and ΔΔe) (Young, 2006).
- The models are continuous density HMMs. All the phone models are 3-state and 16 mixtures. The leading and ending silence/noise model is 3-state and 16 mixtures. The inter-word silence model is 1-state and 16 mixtures.
- Only mono-phone models are used in our preliminary research. We want to use this simple acoustic mapping to identify the basic acoustic and language features before moving to advanced acoustic modeling (e.g., context dependent models) in the next step.

4.2.2 Acoustic model bootstrap

In order to use the wide band recorded WPA speech data to bootstrap a 8 KHz narrow band speech recognition system for mobile voice communication applications, we down sampled the 22050 Hz training data to 8 KHz sampling, along with a 300-3400 Hz band pass filtering to approximate the characteristics of the typical mobile voice communication channels. Along with the WPA speech corpus, the WPA authors provided a pre-built HTK ASR HMM model from WPA speech data to make it easy for us to segment the speech data automatically to the phoneme level. Using the down sampled data, the following two steps shown in Figure 2 are applied to build our bootstrap MSA ASR:

- The segmental K-means and maximum likelihood HMM training algorithms were used for WPA 8 KHz model bootstrap, followed by multiple iterations of Baum- Welch model parameter re-estimation to refine the model until it converged to the best performance by testing it on WPA test set, which produced the WPA 8 KHz ASR system WP_8K_0.

- The WPA to SAMPA phone mapping was applied to WP_8K_0 to create WP_8K_1, which uses SAMPA phoneme set.

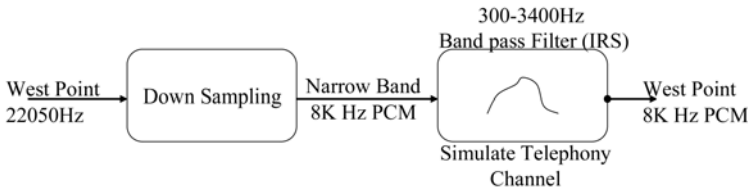

Fig. 2. West Point speech down sampling

4.2.3 SAMPA to WPA phone set mapping

After analyzing of the WPA database, we concluded that the WPA speech data was sufficient for the creation of all the MSA allophone models that can be mapped to the target OrienTel MSA speech recognition using SAMPA phoneme set. Table 5 provides a mapping between them. Since two SAMPA Arabic phones "l'" and "u:" are missing in the WPA system, we used two close phones "l" and "u" to clone them for bootstrap. Then the phone models are rebuilt on the whole Orientel Tunisia MSA training set. Two foreign language SAMPA phones "p" and "V" are not presented in either WPA data or in OrienTel Tunisian database. We fill this gap by borrowing the phoneme models from an English ASR system. Besides, three vowels "aw", "ay", and "ey" appeared in WPA are not used in SAMPA.

4.2.4 Tunisian MSA speech recognition system from WPA recognition system

Once we built the WP_8K_1 ASR system, we experiment both model re-estimation and global MLLR model adaptation to produce MSA ASR systems using OrienTel Tunisian training data. The four MSA speech recognition systems built are:

- WP_8K_1: The baseline telephone SI-ASR system by WPA down sampling.
- BW_6: The retrained on Tunisian MSA training set, with 6 iterations of Baum- Welch parameter re-estimation, bootstrapped by WP_8K_1 model.
- MLLR_1: MLLR model mean adaptation on WP_8K_1, using Tunisian MSA training set.
- MAP_1: MAP adaptation on MLLR_1 on both mean and variances, using Tunisian MSA training set.

Since we are using supervised training in our experiment, we only select a subset of speech recording with good labelling. We also excluded incomplete recorded speech and other exceptions. The total number of retraining and adaptation utterances used in our experiment is 7554.

We test the MSA systems described above by using selected classes of Tunisian MSA data. The test result is shown in Table 6. All the tests are on complete OrienTel Tunisian test set with some excluded data classes described above. We use one pass Viterbi search algorithm with bean control to reduce search space to improve recognition speed. For language model, we use a manual written context free grammar to cover all the content classes described in Table 3, exclude class A, W, S, X not included in test. We also excluded imcomplete recorded and improper labeled speech in Tunisian data collection test set. The total number of utterances we used in test is 1899.

From Table 6, we observed the following

- Acoustic and speaker mismatch can cause significant SI-ASR performance degradation when we use WPA down sampled system (WP_8K_1) to do test on OrienTel Tunisian MSA test set.
- Re-trained OrienTel Tunisian achieves the best performance at combined error rate reduction of 67%, compare to the original system WP_8K_1.
- A simple MLLR global adaptation on miss matched models can achieve 50% error reduction combined, given there are sufficient adaptation data.
- Further MAP adaptation achieves another 2.1% error reduction from MLLR_1 system. It is obvious that MLLR can achieve such good result is due to the main resource of errors from the initial WP_8K_1 is the acoustic channel miss match.

We observed poor results on yes/no (class Q) recognitions. And MAP even make it worse, oppose to overall improvement compare to other classes. From our initial analysis, this is due to pronunciation variations and poorly formed a prior density estimation described in section 2.

Tunisian	MSA	ASR	Systems	
Contents Code	WP_8K_1	BW_6	MLLR_1	MAP_1
I,B	31.25	7.18	9.72	10.19
C	34.13	6.34	14.58	10.84
N	7.33	2.81	3.12	3.64
M	15.42	4.48	7.71	9.95
Q	92.05	37.78	35.56	70.00
D	15.63	9.09	10.39	11.31
T	23.32	11.66	14.13	15.55
E	25.38	12.69	16.24	9.64
O	59.81	25.69	43.58	43.06
L	45.17	18.40	28.35	26.81
Total Error rate	28.01	9.27	14.24	13.94

Table 6. Word error rates (%) on Tunisian MSA ASR systems, bootstrap from WPA ASR

4.2.5 Tunisian to Jordan dialect adaptation

In this section, we experiment adaptation from Tunisian dialect to Jordan dialect. To establish a baseline for comparison, we use the best model built on Tunisian training data, the six iteration of embedded Baum-Welsh trained model BM_6 described in previous section. We name it T1_BW_6 in this section.

Tunisian	MSA	ASR	Systems
ASR Systems	T1_BW_6	G_MLLR	MAP
Error rate	16.85	15.65	16.75

Table 7. Word error rates (%) on Jordan MSA ASR systems, adapted from Tunisian ASR

From Orientel Jordan data collection, we selected 1275 utterances from its test set which constitute about 80 speakers. Table 7 listed our testing results. From column 1, we observed that dialect mismatch degraded ASR performance. Instead of retrain the ASR system, we

randomly selected 600 utterances from Jordan training set to adapt a Tunisian ASR system to Jordan ASR system. This is about 2.5% of total Jordan training set of 23,289 utterances. Using global MLLR adaptation, we saw ASR accuracy improves by 7%. Furthermore, we tested MAP adaptation using the same adaptation set on T1_BW_6 model; we saw slightly accuracy improvement of less than 1%. We believe that for MAP adaptation, more data are needed since it adapts more parameters than produce a global transformation in MLLR.
In this experiment, we used the same recognition algorithm and slightly modified lexicon and context-free grammar as the last section experiment. The lexicon change is based on pronunciation we observed between Tunis and Jordan data collection. Using a PC with 2.4 Ghz Intel processor (Core 2 quad core, but only one is used since our software only use one thread), the 600 utterances adaptation only takes less than 10 minutes.

5. Conclusion

In this chapter, we studied several approaches in building Arabic speaker independent speech recognition for real-world communication service applications. In order to find out practical and efficient methods to build search a system using limited data resource, we study both traditional acoustic model re-estimations algorithms and adaptation methods, which require much less data to improve SI-ASR performance from an existing SI-ASR system with dialect mismatch. Also adaptation methods are more practical to implement as online system to improve SI-ASR at runtime, without restart the system. This is an important feature required by communication service applications, since we need high availability and a little room for down time.
In this work, we only study acoustic model re-estimation and adaptation aspects to improve SI-ASR in mismatched dialect environment. We also observed that there are significant pronunciation variations in different Arabic dialects that need lexicon changes to improve SI-ASR performance. We made lexicon modification when we experiment Tunisia to Jordan dialect adaptation as described above. Also we realize that there are language model variations between different dialects as well.

6. Acknowledgements

The authors wish to thank Col. Stephen A. LaRocca, Mr. Rajaa Chouairi, and Mr. John J. Morgan for their help and fruitful discussion on WPA database for Arabic speech recognition and provide us a bootstrap HMM used in our research.
This work is partially funded by European Commission, under Orientel as an R&D project under the 5th Framework Programme (Contract IST-2000-28373).

7. References

Afify, M., Sarikaya, R Kuo, ., H-K. J., Besacier, L., and Gao Y-Q. (2006). On the use of morphological analysis for dialectal Arabic speech recognition, ICSLP 2006.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains, Ann Math Stat. 37, 1554-1563.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bull. Amer. Math. Soc. 73, 360-363.

Baum, L. E., Petrie, T., Soules, G. and Weiss N. (1970). A Maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Statist. Volume 41, Number 1, 164-171.

Billa, J., Noamany, M., Srivastava, A., Makhoul, J., and Kubala, F. (2002). Arabic speech and text in TIDES OnTAP, Proceedings of HLT 2002, 2nd International Conference on Human Language Technology Research, San Francisco.

Chou, W., Lee, C.-H., Juang, B.-H., and Soong, F. K. (1994) A minimum error rate pattern recognition approach to speech recognition, Journal of Pattern Recognition, Vol. 8, No. 1, pp. 5-31.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, 1977.

Diehl, F., Gales, M.J.F. Tomalin, , M., and Woodland, P.C. (2008). PHONETIC PRONUNCIATIONS FOR ARABIC SPEECH-TO-TEXT SYSTEMS, ICASSP 2008.

Ephraim, Y.,and Rabiner, L. R. (1990) On the relations between modeling approaches for speech recognition, IEEE Trans. on Information Theory 36(2): 372-380.

Hartigan, J. A. and Wong, M. A. (1979). A K-Means Clustering Algorithm, Applied Statistics 28 (1): 100−108.

Huang, X.-D., Acero, A., and Hon, H.-W. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall.

Gauvain, L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. on Speech and Audio Proc., Vol. 2, No. 2, pp. 291-298.

Kirchhoff, K. Vergyri, D. (2004). Cross-dialectal acoustic data sharing for Arabic speech recognition, ICASSP 04.

Kirchhof, K., et. al. (2003). NOVEL APPROACHES TO ARABIC SPEECH RECOGNITION: REPORT FROM THE 2002 JOHNS-HOPKINS SUMMER WORKSHOP, ICASSP 2003.

Jelinek, F. and Mercer, R.L. (1980) Interpolated estimation of Markov source parameters from sparse data, in Pattern Recognition and Practice, Gelsma, E.S. and Kanal, L.N. Eds. Amsterdam, The Netherlands: North Holland, 381-397.

Juang, B.− H., (1985). Maximum likelihood estimation for mixture multivariate observations of Markov chains, AT&T Technical Journal.

Juang, B.− H., Chou, W. and Lee, C.-H. (1996). Statistical and Discriminative Methods for Speech Recognition, in *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, pp 83-108.

Lee, C.-H. and Gauvain, J. L. (1996). Bayesian adaptive learning and MAP estimation of HMM, in *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, pp 83-108.

Lee, C.-H. and Huo Q. (2000). On Adaptive decision rules and decision parameter adaptation for automatic speech recognition, Proceedings of the IEEE, Vol. 88, No. 8, pp. 1241-1269.

Lee, C.-H., Gauvain, J.-L., Pieraccini, R., and Rabiner, L. R. (1993) Large vocabulary speech recognition using subword units, Speech Communication, Vol. 13, Nos. 3-4, pp. 263-280.

LaRocca, S. A., Chouairi, R. (2002). West point Arabic speech corpus, LDC2002S02, Linguistic Data Consortium, http://www.ldc.upenn.edu.

Leggetter C. J. and Woodland P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech & Language, vol. 9, no. 2, 171-185 (15).

MacQueen, J. B., (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Normandin, Y. (1996). Maximum Mutual Information Estimation of Hidden Markov Models, in *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, pp 57-82.

OrienTel, (2001) Multilingual access to interactive communication services for the Mediterranean & the Middle East, http://www.speechdat.org/ORIENTEL/index.html

Rabiner, L. R (1989). A Tutorial on hidden Markov models and selected applications in speech recognition, PROCEEDINGS OF THE IEEE, VOL. 77, NO. 2.

Rabiner, L. R., Juang, B.-H. (1993). Fundamental of Speech Recognition, Prentice Hall, Englewood Cliffs.

Rabiner, L. R., Juang, B.-H., and Lee, C.-H. (1996) An overview of automatic speech Recognition, in *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, pp 1-30.

Rabiner, L. R., Wilpon, J. G., Juang, B.-H. (1986). A Segmental k-means training procedure for connected word recognition, AT&T technical journal 65:33, 21-31.

Rahim, M and Juang, B.-H., (1994) Signal bias removal for robust speech recognition, Proceedings ICASSP-94, Adelaide, Australia.

Rambow, O., et. al. (2006). Parsing Arabic dialects, final report version 1, Johns Hopkins Summer Workshop 2005.

SAMPA, (2005). Speech Assessment Methods Phonetic Alphabet): Computer readable phonetic alphabet, http://www.phon.ucl.ac.uk/home/sampa/home.htm

Schultz, T., and Black, A. W. (2006). Challenges with rapid adaptation of speech translation systems to new language pairs. ICASSP 2006.

Siemund, R. et. al. (2002). OrienTel − Arabic speech resource for the IT Market, LREC 2002 Arabic Workshop.

Wang, Z., Schultz, T., and Waibel, A. (2003). COMPARISON OF ACOUSTIC MODEL ADAPTATION TECHNIQUES ON NON-NATIVE SPEECH, ICASSP 2003.

Young, S. et. al. (2006). The HTK Book, http://htk.eng.cam.ac.uk.

Yu, K. , Gales, M.J.F., and Woodland, P.C. (2008). UNSUPERVISED DISCRIMINATIVE ADAPTATION USING DISCRIMINATIVE MAPPING TRANSFORMS, ICASSP 2008.

Zitouni, I. et. al. (2002). OrienTel: speech-based interactive communication applications for the mediterranean and the Middle East, ICSLP 2002.

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9
Hard cover, 550 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Qiru Zhou and Imed Zitouni (2008). Arabic Dialectical Speech Recognition in Mobile Communication Services, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from: http://www.intechopen.com/books/speech_recognition/arabic_dialectical_speech_recognition_in_mobile_communication_services

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821