Daniel Régis Sarmento Caon

Automatic speech recognition, with large vocabulary, robustness, independence of speaker and multilingual processing

Vitória - ES, Brasil

27 de agosto de 2010

Daniel Régis Sarmento Caon

Automatic speech recognition, with large vocabulary, robustness, independence of speaker and multilingual processing

Dissertação apresentada para obtenção do Grau de Mestre em Informática pela Universidade Federal do Espírito Santo.

Orientador:
Thomas Walter Rauber

Co-orientador:
Rodrigo Varejão Andreão

DEPARTAMENTO DE INFORMÁTICA
CENTRO TECNOLÓGICO
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Vitória - ES, Brasil

27 de agosto de 2010

Dissertação de Mestrado sob o título *"Automatic speech recognition, with large vocabulary, robustness, independence of speaker and multilingual processing"*, defendida por *Daniel Régis Sarmento Caon* e aprovada em 27 de agosto de 2010, em Vitória, Estado do Espírito Santo, pela banca examinadora constituída pelos professores:

Prof. Dr. Thomas Walter Rauber
Orientador

Prof. Dr. Rodrigo Varejão Andreão
Instituto Federal do Espírito Santo

Prof. Dr. Flávio Miguel Varejão
Universidade Federal do Espírito Santo

Prof. Dr. Carlos Alberto Ynoguti
Instituto Nacional de Telecomunicações

Resumo

Este trabalho visa prover assistência cognitiva automática via interface de fala, à idosos que moram sozinhos, em situação de risco.

Expressões de angústia e comandos vocais fazem parte do vocabulário alvo de reconhecimento de fala.

Durante todo o trabalho, o sistema de reconhecimento de fala contínua de grande vocabulário Julius é utilizado em conjunto com o Hidden Markov Model Toolkit(HTK). O sistema Julius tem suas principais características descritas, tendo inclusive sido modificado. Tal modificação é parte da contribuição desse estudo, assim como a detecção de expressões de angústia (situações de fala que caracterizam emergência).

Quatro diferentes linguas foram previstas como alvo de reconhecimento: Francês, Holandês, Espanhol e Inglês. Nessa mesma ordem de linguas (determinadas pela disponibilidade de dados e local de cenários de integração de sistemas) os estudos teóricos e experimentos foram conduzidos para suprir a necessidade de trabalhar com cada nova configuração. Este trabalho inclui estudos feitos com as linguas Francês e Holandês.

Experimentos iniciais (em Francês) foram feitos com adaptação de modelos ocultos de Markov e analisados por validação cruzada.

Para realizar uma nova demonstração em Holandês, modelos acústicos e de linguagem foram construídos e o sistema foi integrado a outros módulos auxiliares (como o detector de atividades vocais e sistema de diálogo).

Resultados de reconhecimento de fala após adaptação dos modelos acústicos à um locutor específico (e da criação de modelos de linguagem específicos para um cenário de demonstração do sistema) demonstraram 86,39% de taxa de acerto de sentença para os modelos acústicos holandeses. Os mesmos dados demonstram 94,44% de taxa de acerto semântico de sentença.

Palavras-chave: Processamento de Sinais de Fala; HTK; Julius; K-Fold; Modelos Ocultos de Markov; Modelagem Acústica.

Abstract

This work aims to provide automatic cognitive assistance via speech interface, to the elderly who live alone, at risk situation.

Distress expressions and voice commands are part of the target vocabulary for speech recognition.

Throughout the work, the large vocabulary continuous speech recognition system *Julius* is used in conjunction with the *Hidden Markov Model Toolkit*(HTK). The system *Julius* has its main features described, including its modification. This modification is part of the contribution which is in this work, including the detection of distress expressions (situations of speech which suggest emergency).

Four different languages were provided as target for recognition: French, Dutch, Spanish and English. In this same sequence of languages (determined by data availability and the local of scenarios for the integration of systems) theoretical studies and experiments were conducted to solve the need of working with each new configuration. This work includes studies of the French and Dutch languages.

Initial experiments (in French) were made with adaptation of hidden Markov models and were analyzed by cross validation.

In order to perform a new demonstration in Dutch, acoustic and language models were built and the system was integrated with other auxiliary modules (such as voice activity detector and the dialogue system).

Results of speech recognition after acoustic adaptation to a specific speaker (and the creation of language models for a specific scenario to demonstrate the system) showed 86.39 % accuracy rate of sentence for the Dutch acoustic models. The same data shows 94.44 % semantical accuracy rate of sentence.

Keywords: Automatic speech recognition; HTK, Julius; K-Fold; Hidden Markov Models; Acoustic modeling.

Dedicatória

Dedico este trabalho aos honestos e justos.

Contents

List of Figures

List of Tables

"*O otimista é um tolo.*
O pessimista, um chato.
Bom mesmo é ser um realista esperançoso."
Ariano Suassuna

1 Introduction

In this chapter, the motivation of this work is initially explained at the section 1.1. The requirements are listed in section 1.2, followed by the objectives (section 1.3). Finally, the organization of this work comes at section 1.4.

1.1 Motivation

This research is motivated by the increasing elderly population who is living alone in Europe, who daily need assistance and health care [Clement, Tennant e Muwanga 2010]. As the number of persons suffering from cognitive disabilities in Europe increases because of the increasing number of elderly people and pathologies associated with aging such as Alzheimer's disease or depression, there is a social and economical pressure for staying at home as long as possible.

Life monitoring systems, composed by many kinds of sensors (infra-red, sound, pulse etc) and artificial intelligence have great potential to save lives [Baldinger et al. 2004].

The use of microphones can give interesting feedback of situation awareness. It is complementary to the infra-red sensors, which in a specific danger situation, for example, may not be able to localize the care recipient (CR, is the elderly person who lives alone and being monitored) laying down behind the furniture and calling for help.

Not only the simple command words can be recognized to make life easier for the CR, but distress expressions like "Help me!" can quickly make the smart house or the companion robot call an ambulance and the doctor.

Moreover, the integration of a dialogue system permits the creation of cognitive assistance. The automatic cognitive exercises are very important to maintain the mental health of the elderly population.

1.2 Requirements

To build a reliable speech recognition system, the acoustic models and language models have to be adapted to specific situations and speakers (e.g. [Goronzy e Kompe 1999]). Normally, speech recognition systems work with a close-talk microphone, inside a car or in front of a computer, but in this situation, almost all the house has to be listened. The use of accessories like carpets or curtains is recommended in order to avoid reverberation (specially in case there are large windows). Some more robustness is also desired to distinguish speech from some detected noise like laundry machines or keys falling down (Voice Activity Detector). Moreover, a top-level dialog system is capable to add semantic comprehension to the situation and increase the accuracy of the speech signal recognition. In order to link all the chain of sub-modules concerning sound signal acquisition, voice activity detection, speech recognition and dialog system, an integration campaign is done through the use of a system awareness black-board via SOAP (Simple Object Access Protocol) connections.

1.3 Objective

This work has been developed in the context of an European Project[1].

The objective is mainly to provide an integrated cognitive assistance. The care recipient (elderly person being monitored) live with the companion robotic systems for ability and for security. The speech recognition system can be integrated in the house system and in the companion robotic system for detecting commands and also detecting distress expressions spoken in dangerous situations.

The figure 1.1 illustrates a situation where the care recipient is receiving full time attention.

In this project, the speech recognition system must recognize continuous speech in different languages(including French, Spanish, Dutch and English), work 24 hours a day and be in real time. Noises generated in the environment, like chairs moving, water flowing and objects falling down may be interfering on the desired recognition targets. Thus, the choice of the recognition engine is an important first step.

The multi-language speech recognition accuracy depends on the robustness of the system. The perfect engine should use the state of the art algorithms. There are different recognition engines available today, in the state of the art and some of them are also free for research

[1]This work has been conducted within the FP7(Seventh Framework Program) Integrated European Project CompanionAble (http://www.companionable.net).

Figure 1.1: Ilustrative example of a care recipient being monitored.

purposes.

One module of the monitoring system consists in analyzing the speech flow of the person living in the monitored room in order to recognize not only noises but also distress expressions. To do this, we decided to use an Automatic Speech Recognition (ASR) system based on the open-source software Julius([Lee, Kawahara e Shikano 2001]). Recent developments of Julius, describing how it incorporates major state-of-the-art speech recognition techniques have been recently published [Lee e Kawahara 2009].

This recognition engine is a state-of-the-art large vocabulary and continuous speech ASR engine. It can be configured to make real-time recognition and uses standard acoustic Hidden Markov modelization with stochastic n-gram or deterministic grammar. Moreover this engine is a good time-memory-performance trade-off.

To define our ASR system we have to setup acoustic and linguistic models. In a first step we focus on French but other language like Spanish, Dutch and English must also be considered for the project. To develop our ASR system, we made a preliminary study on three French distress expression databases.

Summarizing, the aim of this work is to propose and validate an automatic speech recognition system with the following characteristics:

- Online speech acquisition, processing and decoding;

- HMM adaptation to a target speaker;
- Validation of the system using databases in two different languages (French and Dutch).

An important contribution of the work is the fact that it has evaluated databases with an uncommon type of speech signal (distress expressions, spoken in French and Dutch languages).

1.4 Organization of this work

This work is organized as follows: Chapter 1 introduces the works, defining the initial proposition of this research, the motivation and the concerned objectives. Chapter 2 describes architecture of an automatic speech recognition system. Chapter 3 provides details about the speech databases identified as useful for this project and directly related to the acoustic modeling (chapter 4). The main implementations and experiments (specially the adaptation experiments with k-fold cross validation) conducted within this work are described in chapter 5. Chapter 6 describes work missions and the integration with support systems towards the Dutch demonstration of integrated systems. The global conclusion and perspectives come in chapter 7. Finally, the acknowledgment is provided in chapter 8, followed by the appendix (which mainly contains information about database acquisitions).

2 *Architecture of an Automatic Speech Recognition System*

2.1 Introduction

The main concepts concerning the architecture of an Automatic Speech Recognition (ASR) system are described in this chapter. First, the basic ASR concepts are explained. The architecture (concerning the configuration of Julius and HTK systems) that was used in this work is shown.

Further, the information about the language modeling, feature extraction and decoding configurations are described with more details.

Finally, systems that are considered the state of the art are described.

2.2 Speech Recognition Basics

A speech recognition system can be divided into two parts: Pre-processing and decoding.

The pre-processing part concerns mainly the feature extraction. This step depends on the choice of the data modelization.

The data modelization is the manner that the data (in this case, the speech signal) will have its main features extracted and represented into the hidden Markov models. The speech signal is not directly used and some filters and transformations are applied in the Feature Extraction step to provide data related to speech (in the most possible discriminative and compressed representation to the decoder). The feature extraction is based on the extraction of Mel-Frequency Cepstral Coefficients (MFCC). It will be described at section 2.3.

The speech signal can be analyzed by continuous or discrete models. In the continuous case, the speech modeling is usually done by continuous hidden Markov models using probability density functions (usually Gaussian mixtures are used to get better results). In the discrete case,

there is an additional step named Vectorial Quantization [Gersho e Gray 1992] , used after the feature extraction. The Vectorial Quantization provides discrete parameters to build the discrete hidden Markov models. The discrete model processing is faster than the continuous one, but the Vector Quantization naturally provides some loss of signal information.

Moreover, the memory storage capacity and processing capabilities are not a problem as years ago. Therefore, the hidden Markov Models in this work are continuous to get the best recognition accuracy instead of faster signal processing time.

The decoding part concerns the recognition engine (Julius) based in hidden Markov models. It is a two-step speech recognition system which uses basically Viterbi [Rabiner e Juang 1993] at the first pass and the A* (called "A Star") algorithm at the second pass. More details are described at section 2.4.

The global schema of an Automatic Speech Recognition system (ASR) is presented in the figure 2.1.

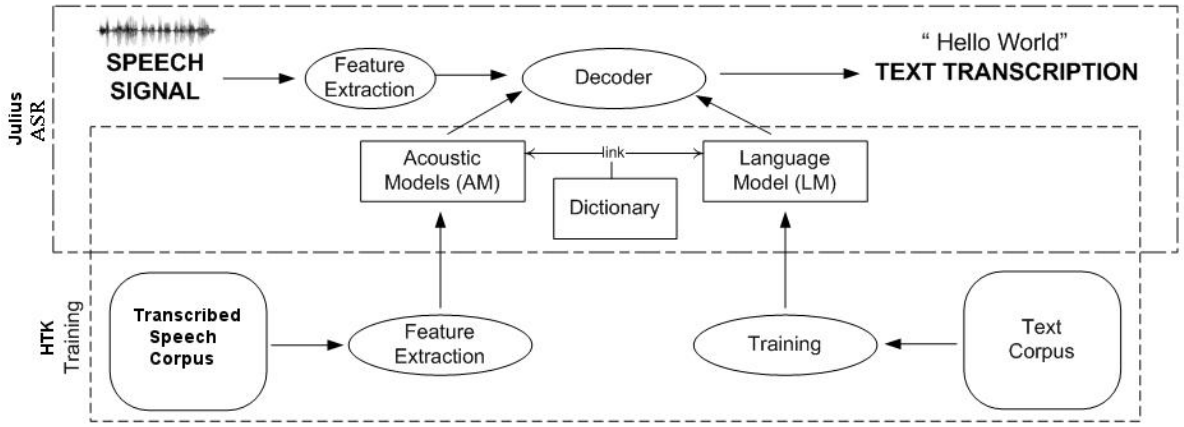

Figure 2.1: Architecture of an ASR system based on statistical approaches

As it is possible to observe at figure 2.1, a standard Automatic Speech Recognition system has as input the speech signal wave and as output the transcribed sentence related to that signal. The Julius system is used within the recognition steps, while the HTK is applied within the training steps. Initially, the possible language words should be known, and the possible sounds should be related to each word and acoustically modeled. The feature extraction (which concerns the computation of the Mel-Frequency Cepstral Coefficients) is described at section 2.3. The dictionary (also called lexicon, described at 2.4.4.1) makes the link between sounds and words.

2.3 Computing Mel-Frequency Cepstral Coefficients

One of the most widely used **feature extraction** schemes is the computation of the MFCC (see [Benesty, Sondhi e Huang 2008, Davis e Mermelstein 1980]). Thus, the MFCC parametrization is also used in this work.

The MFCC Feature Extraction employs in each signal frame usually some main steps: pre-emphasis filtering, applying a sliding Hamming window and extracting Mel-frequency cepstral coefficients from the signal frames. A signal frame is an interval of the speech signal (in this work, it has 32ms of duration) where the feature extraction (illustrated at figure 2.2) is done. The complete process is described below.

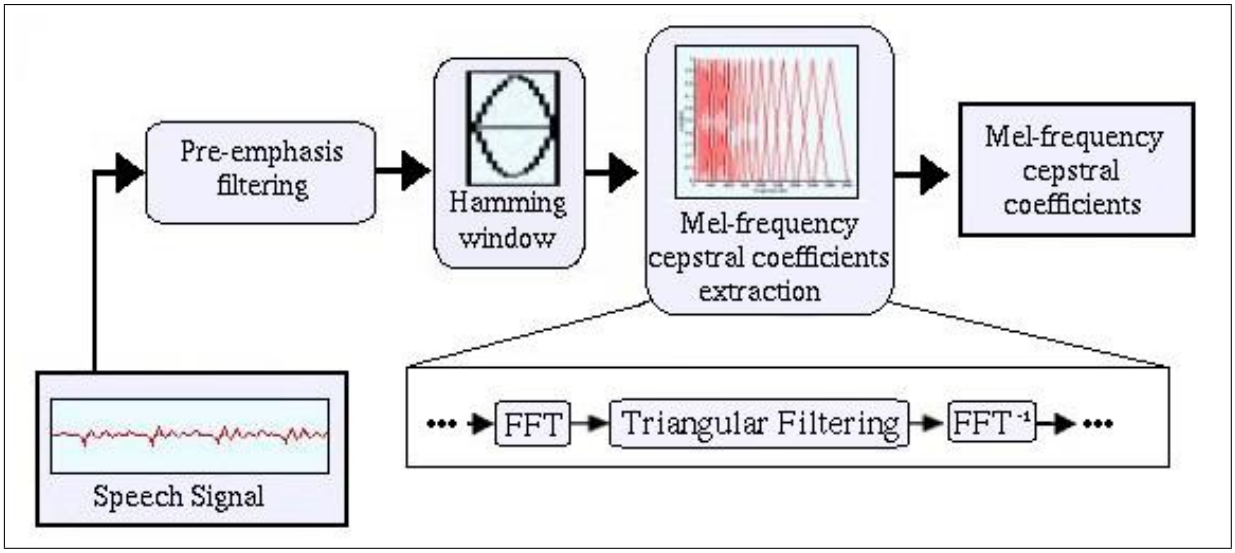

Figure 2.2: MFCC Feature Extraction

The pre-emphasis filtering has the function to equalize the speech signal, compensating the high frequencies attenuation (which usually occurs on glotal spectrum)[Picone93].

The sliding Hamming window will focus the next MFCC feature extraction step in a specific signal frame. In this work the window is 32 ms long and after each extraction it slides 10 ms further. When applied, the Hamming window multiplies the frame samples, smoothing the signal "borders" and improving the spectral representation.

The Hamming window formula is shown below:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N-1} & \text{if } 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The quasi-stationary signal is the desirable input for the Fast Fourier Transform (FFT). The

FFT provides a magnitude frequency response for each frame.

After the FFT, a triangular band-pass filtering is applied with a set of 20 Mel frequency triangular filters (figure 2.3) in order to get the log energy of each filter. The Mel-frequency scale is used to simulate the human listening capability.

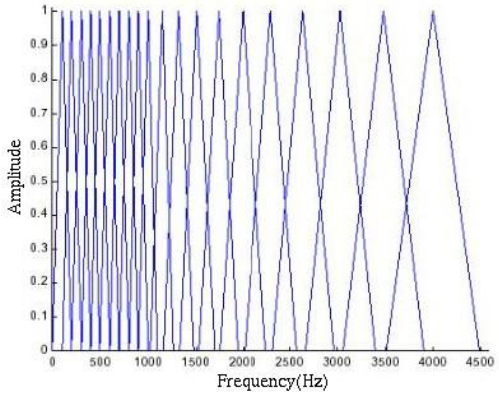

Figure 2.3: The 20 triangular filters of the Mel filter bank

The general Mel-frequency formula is

$$Mel(f) = 2595 \log_{10}\left(\frac{f}{700}+1\right) = 1127 \log_{e}\left(\frac{f}{700}+1\right), \tag{2.2}$$

where f is the frequency in Hz.

The Inverse Fast Fourier Transform (FFT^{-1}) uses the 20 log energy measures to provide the mel-scale cepstral coefficients.

For better performance, Log Energy and Delta Cepstrum operations can be used. The Log Energy uses the energy within a frame and is usually a 13^{rd} feature to MFCC. Pitch, zero cross rate, high-order spectrum momentum, and other features can also be extracted.

Time derivates from the energy and also from the MFCC can be useful new features. Most ASR systems (including this work) use 12 MFCC, energy and their first and second derivates, hence 39-dimensional features are built.

2.4 Decoding

The principles of a recognition engine based in hidden Markov models are explained in this section, describing how the Large Vocabulary Continuous Speech Recognition (LVCSR) engine Julius works.

After MFCC feature extraction, the recognition engine has as input a sequence O of T observation vectors, $O = (o_1, o_2, ..., o_T)$,where o_i has size 39. Recognizing a speech utterance means obtaining the sequence of n words $W^* = w_1 ... w_n$ which maximizes the probability of the words given the observation sequence O. This problem can be written as below:

$$W^* = \underset{W}{\operatorname{argmax}} P(W|O) \tag{2.3}$$

However, direct computing of $P(W|O)$ is hard or even impossible ([Rabiner 1989, Razik 2007]). Thus the Bayes theorem is used to reformulate the last equation as:

$$W^* = \underset{W}{\operatorname{argmax}} \frac{P(O|W)P(W)}{P(O)} \tag{2.4}$$

Now the problem can be expressed in function of three other probabilities:

P(O|W) Acoustic probability. The probability to observe the observation sequence O given the sequence of words W.

P(W) Linguistic probability. The probability *a priori* for the sequence of words W.

P(O) The observation's probability.

Given a observation sequence O, $P(O)$ does not depend on the studied sequence of words W. The equation 2.4 is then upgraded to a new equation 2.5 which is composed only by the acoustic and linguistic probabilities:

$$W^* = \underset{W}{\operatorname{argmax}} P(O|W)P(W) \tag{2.5}$$

In order to solve this equation, the acoustic probabilities $P(O|W)$ and linguistic probabilities $P(W)$ have to be estimated for all possible sequence of words W, finally obtaining the maximum score $P(O|W)P(W)$.

Considering now the sequence W only limited by hidden Markov models M of a word, then $P(O|W) = P(O|M)$. To obtain $P(O|M)$ directly following all possible sequences of states leads to a problem of combinatorial explosion. In fact, for a hidden Markov model of N states, the complexity of this calculation is $O(T.N^T)$, being T the length of the observation sequence O. Therefore, we need to use an algorithm for calculating this with a more reasonable complexity. Usually the Viterbi algorithm (or its variations) is used, which has a complexity of $O(T.N^2)$.

2.4.1 The Viterbi Algorithm

2.4.1.1 The Viterbi Principle

The objective is to find the sequence of states that maximizes $P(O|M,Q)$. The problem to solve can be shown in a graphic with 2-axis coordinates (figure 2.4). The sequence of observations is in the abscissa and the model M in the ordinate.

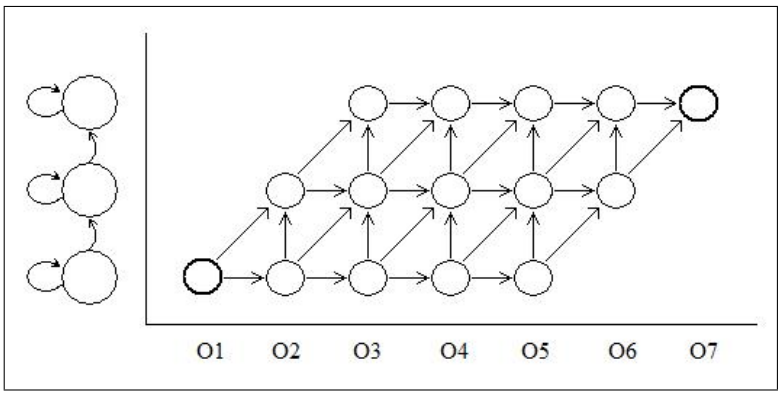

Figure 2.4: Viterbi Graph of a left-right HMM with 3 states, for 7 observations

In this graph (figure 2.4), one node represents one state i of the model for one specific observation o_t with an associated value whose probability is given by $b_i(o_t)$. The arcs are transitions from one state i to a state j (i can be equal j) whose transition probability is given by a_{ij}.

The solution will be represented then by one path which starts at the state of the HMM at the frame $t = 1$ and ends at the extremity the state of the model at the frame $o = T$.

In the figure 2.4, a left-right HMM is used. The left-right is an useful topology which restricts many of the possible paths. This topology, in fact, forces the path to start in the first state of the model and to end in its last state.

The Viterbi Algorithm permits to search the best path by using the Bellman's Principle of Optimality.

2.4.1.2 Bellman's Principle of Optimality

The Bellman's principle of optimality is used in dynamic programming.

It can be applied to the path search of a graph and it is expressed as the following manner:

Considering that we know the optimal path to arrive at A_1, A_2 and A_3 (figure 2.5).

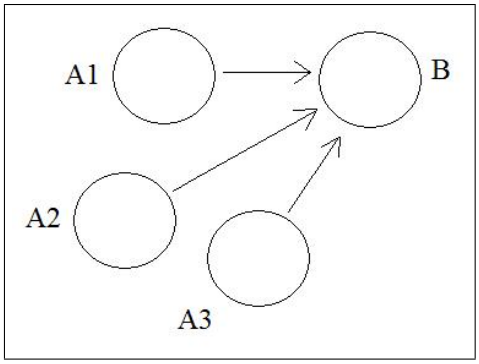

Figure 2.5: Example of a graph part to explain the Bellman's principle of optimality

Thus, the optimal path to go to the node B, the variable $d\_acumul(B)$, is given by the equation 2.6

$$d\_acumul(B) = \min \begin{cases} d\_acumul(A_1) + d(A_1, B) \\ d\_acumul(A_2) + d(A_2, B) \\ d\_acumul(A_3) + d(A_3, B) \end{cases}, \tag{2.6}$$

where $d\_acumul(A_i)$ is the accumulated distance to arrive at A_i, and $d(A_i, B)$ is the local distance to go from A_i to B, $\forall i \in \{1, 3\}$

2.4.1.3 The Viterbi Algorithm

Given the figure 2.4 and the Bellman's principle of optimality, it is possible to assume that the accumulated scores for each state of the hidden Markov model in the frame t depend only on:

- The accumulated score to each state before the frame $t-1$,
- the transition probabilities between the states at the frame $t-1$ and the state at the frame t.
- the probabilities to emit the observation o_t for the states of the HMM.

Being $\delta_t(j)$ the probability of the best path which stops at the frame t at the state j of the HMM, the following relation (equation 2.7) is obtained:

$$\delta_t(j) = \max_i \left(\delta_{t-1}(i) \times a_{ij} \times b_j(o_t) \right) \tag{2.7}$$

The Viterbi algorithm is based in this recurrence relation. At each frame t it is possible to calculate a new probability for the best path to reach the state j of the model. This way, all the possible paths are evaluated, and the best score confirms the best path. The initial problem (to calculate $P(O|M)$) can be linked to the equation 2.7 by the relation 2.8:

$$P(O,Q/M) = \max_i \delta_T(i) \tag{2.8}$$

The main steps of the Viterbi algorithm can be described as:

Initialization: Set $\delta_0(i) = \pi_i$, the initial probability of being at one of the states of the hidden Markov model.

Recurrence: At the frame t, for each state i of the model, $\delta_t(i)$ is calculated using the equation 2.7 (which depends only on the δ_{t-1} values).

Termination: For each state i of the model, search for the maximum $\delta_T(i)$ where $1 <= i <= N$. This way, $P(O,Q/M)$ is obtained (see eq. 2.8).

As the algorithm depends on the number of states in the model, and at each frame t it depends only on the accumulated scores at the frame $t-1$, the complexity becomes linear and related to the length of the observation sequence. It is also important to note that, in the algorithm, not only $P(O|M)$ can be estimated, but the best sequence of states Q can also be known. In fact, to know the best state sequence, it is just a matter of going back from the state at frame T to the state at the frame 1 (this is called backtracking procedure).

2.4.1.4 The Viterbi Algorithm in the case of a sentence

In this section, the use of the Viterbi Algorithm was explained for a general case of a hidden Markov Model M to calculate $P(O|M)$, where M is related to the phonetic units of a word. It was shown that using this algorithm, it is possible to determine the best state sequence of the model given the observations sequence of a word. In order to recognize one sentence or a sequence of phonemes, the principle is the same. In fact, a Hidden Markov Meta-Model (a model containing models) has to be built, and each meta-state will represent one lexicon's word (thus, the meta-state of a word contains states of phones). The meta-model is ergodic, all the transitions between the meta-states are possible and depend on the language model. Therefore, the best sequence of meta-states, calculated by the Viterbi Algorithm, will correspond to a sentence (the solution

of the recognition system). To find the sequence of words W^* that maximizes the equation 2.5 means to search to sequence of words that maximize the following value:

$$\max_{W \in \Xi} \pi_{w_0} \prod_{w_1 \in W} P(O|w_i) P(w_i|w_{i-1}...w_0) \tag{2.9}$$

, where Ξ represents the possible sequences of words that can be built using the words from the lexicon, π_{w_0} is the probability of the first word in a sequence W, $P(O|w_i)$ (equivalent to the emission probability) is the acoustic probability of the word w_i of the sequence W, and $P(w_i|w_{i-1}...w_0)$ is the linguistic probability n-gram,which contains the values to the transition probabilities inside the hidden Markov model (the n-grams are described at 2.4.4.2).

Due to the Bellman's principle of optimality, each word in the graph (word net) is the extremity of an unique path starting in the beginning of the sentence. Thus, for each word of the graph, the precedent word by the path is determined in an unique way.

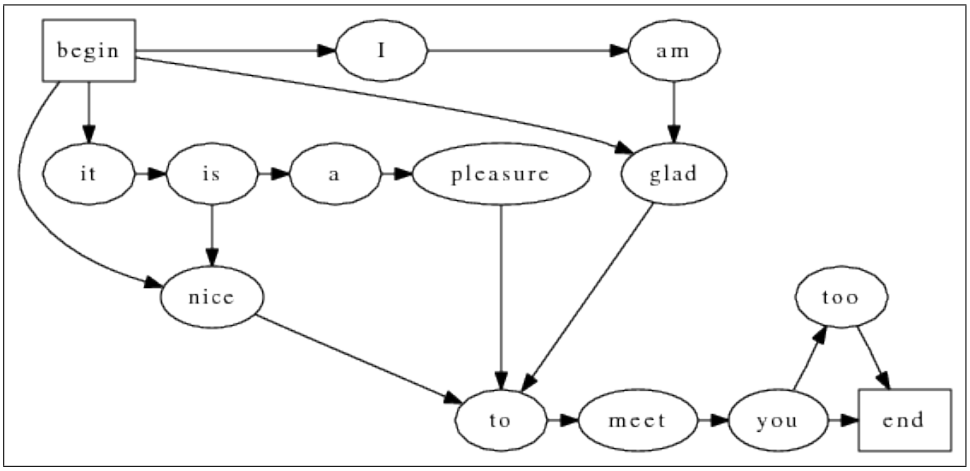

Figure 2.6: Example of a word net.

2.4.2 The A* Algorithm

The A* algorithm is also a very known method to search for the best path in a graph [Russell e Norvig 2002]. This algorithm is commonly used in artificial intelligence (specially games) to find the optimal path between two points.

The LVCSR Julius is a two-step recognition engine, which uses the algorithms Viterbi and A*, for its first and second steps of recognition respectively. The recognition makes use of the stack data structure.

At the second pass (A* algorithm), the search for the best path will continue to be done

by Julius using the word graph built by the first pass (Viterbi algorithm). The second pass will search for the best path in the inverse way of the word sequences (inversed n-grams), reducing the list of opened hypothesis until it reaches the n-best results required by the user.

2.4.3 N-Best Sentences

Every time that an audio input from live microphone recognition or from a file is processed by the Julius engine, the best hypothesis of sentence can be provided (see table 2.1). Not only this, the n-best sentences can also be provided sorted by acoustic and/or language scores. The variable *n* is set by command line at runtime. That is a very good aspect that permits a dialog manager to identify in the n-best sentences the inputs of interest, considering that the speech recognition result is usually not 100% perfect.

Taking advantage on the multiple outputs given by the Julius engine, [Razik 2007] studies new confidence measures to integrate in this system.

sentence 1:	la reconnaissance automatique de_la parole
sentence 2:	avec reconnaissance automatique de_la parole
sentence 3:	la reconnaissance automatique apparent
sentence 4:	la reconnaissance automatique apparents
sentence 5:	à la reconnaissance automatique de_la parole
sentence 6:	la reconnaissance automatiquement par an
sentence 7:	la reconnaissance automatique des parents
sentence 8:	avec un espace automatiquement par an
sentence 9:	avec un espace automatiquement pates
sentence 10:	avec l' espace automatiquement par an

Table 2.1: The 10-best results for an utterance in French

2.4.4 Lexicon and Language Model

2.4.4.1 Lexicon

The speech recognition system needs a list of possible words to recognize. This list is called Lexicon (or Dictionary). It contains the definition of the words, concerning how it is written and all the possible pronunciations. A word which is not in the Lexicon will never be recognized.

One way to build a new Lexicon is to extract from a large text corpus the most frequent words. For the recognition systems called as "Large Vocabulary", the lexicon's size is around 65k words (65535 words is the maximum by default for the Julius system, used in this work).

Another important point when building the lexicon is to pay attention to the possible variations of pronounciations caused when some words are linked.

The table 2.2 shows part of a lexicon for French words.

Word	Output	Pronunciation
...	...	...
multilatérales	[multilatérales]	m y l t i l a t eh R a l swa
multilatéralisme	[multilatéralisme]	m y l t i l a t eh R a l i s m swa
multilatéraux	[multilatéraux]	m y l t i l a t eh R o z_l
multilingue	[multilingue]	m y l t i l in g swa
multimédia	[multimédia]	m y l t i m eh d j a
multimédias	[multimédias]	m y l t i m eh d j a z_l
multinational	[multinational]	m y l t i n a s j oh n a l
multinationale	[multinationale]	m y l t i n a s j oh n a l swa
multinationales	[multinationales]	m y l t i n a s j oh n a l swa
multinationaux	[multinationaux]	m y l t i n a s j oh n o z_l
...	...	...

Table 2.2: Words of a French lexicon

2.4.4.2 Language Model

One important part of a speech recognition system is the language model. The language models also have the objective to represent the behavior of a language. It can determine if a sentence is acceptable or not, considering the modeled language.

- Deterministic Grammars *versus* Sthochastic N-Grams

  The Julius Engine can use language models that are deterministic (using linguistic information contained in grammars) or sthochastic (using statistical information contained in n-grams) [1].

  Grammars

  Usually, the linguistic knowledge is used to build the language model, treating the structure of a sentence to validate the conditions of acceptance. The oral language, however, has a spontaneous aspect which usually requires manual intervention on the construction of a sentence (not always respected). The type of speech (read texts, telephone dialogues, interviews, news, meetings etc) can influence the spoken language in a way not previewed by the linguistic knowledge.

[1] The Julius recognition mode has to be set in compilation time (e.g. the option Julian has to be enabled at the compilation to make recognition using grammars).

For the grammars, the construction process is described at the Julius official website[2].

N-Grams

In the automatic speech recognition systems, the language can be modeled by another category of language model, called statistic models. Those models are learned automatically by the use of very large text corpus (millions of words). The n-gram modelization is usually the best choice, capable to estimate automatically, by the text corpus, the probabilities for sequences containing n words. The frequency for each sequence that appears will set the probabilities for 1-grams (probability of one word to appear), 2-grams (probability of two specific words appearing in sequence) and continuously to n-grams (n being previously set as the maximum number of words in sequence, to estimate the probability of appearing).

The common automatic procedure to build n-grams (forward 3-gram and backward 5-gram) in order to permit the Julius to launch the first and second steps using the statistical language probabilities are described at the Julius forum website [3].

2.5 Automatic Speech Recognition Systems

Contrary to HTK, the Julius and the CMU Sphinx systems use a BSD style license which does not restrict redistribution.

- HTK

  The Hidden Markov Model Toolkit [Young et al. 2006] is currently in version 3.4 (available on the Internet [4]). The HTK was originally developed by the Department of Engineering, University of Cambridge. The HTK is a very flexible system of signal recognition, which can be used for speech synthesis, recognition of characters, for DNA sequencing and other applications. The HTK consists in modules, libraries and tools made in C-language. The tools facilitate the analysis of a large quantity of speech items, HMMs training, testing and analysis of results. The software supports both continuous mixture Gaussian distributions as discrete, being able to create complex systems based on HMMs.

- JULIUS

  "Julius" [Lee e Kawahara 2009] is a high-performance, two-pass large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and

[2] http://julius.sourceforge.jp/en_index.php?q=en_grammar.html
[3] http://julius.sourceforge.jp/forum/viewtopic.php?f=5&t=132
[4] htk.eng.cam.ac.uk/

developers. Based on word N-gram and context-dependent HMM, it can perform almost real-time decoding on most current PCs for 60k word dictation task. Major search techniques are fully incorporated such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning, Gaussian selection, etc. Besides search efficiency, it is also carefully modularized to be independent from model structures. Various HMM types are supported such as shared-state triphones and tied-mixture models, with any number of mixtures, states, or phones. Standard formats are adopted to cope with other free modeling toolkit such as HTK, CMU-Cam SLM toolkit etc. The main platform is Linux and other Unix workstations, and also works on Windows. Most recent version is developed on Linux and Windows (cygwin / mingw), and also has Microsoft SAPI version. Julius is distributed with open license together with source codes. Julius has been developed as a research software for Japanese LVCSR since 1997, and the work was continued under IPA Japanese dictation toolkit project (1997-2000), Continuous Speech Recognition Consortium, Japan (CSRC) (2000-2003) and currently Interactive Speech Technology Consortium (ISTC).

- SPHINX

  Sphinx4 [Walker et al. 2004] is a state-of-the-art speech recognition system written entirely in the JavaTM programming language [5] . It was created via a joint collaboration between the Sphinx group at Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL), and Hewlett Packard (HP), with contributions from the University of California at Santa Cruz (UCSC) and the Massachusetts Institute of Technology (MIT). Sphinx-4 started out as a port of Sphinx-3 to the Java programming language, but evolved into a recognizer designed to be much more flexible than Sphinx-3, thus becoming an excellent platform for speech research. Some of Sphinx capabilities are the live mode and batch mode speech recognizers, capable of recognizing discrete and continuous speech and the generalized pluggable language model architecture, which includes pluggable language model support for ASCII and binary versions of unigram, bigram, trigram, Java Speech API Grammar Format (JSGF), and ARPA-format FST grammars.

[5] http://cmusphinx.sourceforge.net/sphinx4

3 Databases

3.1 Introduction

In this chapter, the databases used in this work are described. Some of them are currently identified as being databases with great potential for further research. Databases are listed here separated by French and Dutch languages.

3.2 Databases in French

3.2.1 Multiple Speakers

3.2.1.1 ESTER

The ESTER Corpus ([Galliano et al. 2006]) database contains French broadcasting news. Around 40 hours were used to train the acoustic models which are employed in the adaptation experiments of section 5.2. The acoustic models trained with the ESTER Corpus represent now the best models available in the French language for this project.

3.2.1.2 Distress Expressions

The Clean French Distress Expressions database, provided by ESIGETEL, has 21 speakers, each one speaking 126 different distress sentences, totalling 2646 distress utterances, e.g. “Je suis tombé!” (“I’ve felt down!”) and “Ne me laissez pas tout seul!” (“Don’t leave me alone!”).

Another similar French Distress Expressions database, also provided by ESIGETEL, with 216 utterances and different SNRs (0, 10, 20 and 40 dB) is available to analyse the noise impact in the recognition (Noisy French database).

A new database mixing the Clean Distress Expressions database with noise recordings (sounds from laundry and bathroom) were built. Each noise is cut in the same length of each

distress expression and mixed together at different SNRs.

3.2.1.3 Phonologie du Français Contemporain (PFC)

The PFC Corpus contains mainly young speakers and was not used in the CompanionAble research. Around 600 hours of French around the world are recorded.

3.2.1.4 La Collégiale v1

The database is recorded with 23 elderly speakers acting in a scenario. The recordings are done using multiple microphone channels, simultaneously and in different positions of the site. Around 1 hour of continuous sound recording is done for each speaker. The site's floorplan is shown at the section B.2.

3.2.2 Single Speaker

3.2.2.1 Speaker Dependent Readings (SDR)

The Speaker Dependent Readings are recorded by one male non-native French speaker. This database contains 162 sentences, most of them are repeated (from 3 to 8 times, usually around 5) for better pronunciation.

3.2.2.2 Speaker Dependent Interviews (SDI)

The Speaker Dependent Interviews are separated into two parts. One part recorded by one male speaker (SDIm) and the other by one female speaker (SDIf). Both are non-native French speakers.

3.3 Databases in Dutch

3.3.1 Groningen

Over 20 hours of speech (read speech material from 238 speakers) are provided. The speakers read:

- 2 short texts (with many quoted sentences to elicit 'emotional' speech)
- 23 short sentences

- 20 numbers (the numbers 0–9 and the tens from 10–100)
- 16 monosyllabic words (containing all possible vowels in Dutch)
- 3 long vowels (a:, E, i)

3.3.2 Spoken Dutch Corpus (CGN)

The version 2 of the Spoken Dutch Corpus (*Corpus Gesproken Nederlands; CGN*) is the larger speech database obtained in Dutch and Flemish (both are dialects from Netherlands).

More important than showing the number of hours recordings is to show the number of spoken words.

The table 3.1 can show details about the data provided in this large Corpus.

Component-Description	Total number of words	Flemish(VL)	Dutch(NL)
a. Spontaneous conversations('face-to-face')	2,626,172	878,383	1,747,789
b. Interviews with teachers of Dutch	565,433	315,554	249,879
c. Spontaneous telephone dialogues	1,232,636	489,100	743,537
d. Spontaneous telephone dialogues (local)	853,371	343,167	510,204
e. Simulated business negotiations	136,461	0	136,461
f. Interviews/discussions/debates (broadcast)	790,269	250,708	539,561
g. (political) Discussions/debates/meetings	360,328	138,819	221,509
h. Lessons recorded in the classroom	405,409	105,436	299,973
i. Live (e.g. sports) commentaries (broadcast)	208,399	78,022	130,377
j. Newsreports/reportages (broadcast)	186,072	95,206	90,866
k. News (broadcast)	368,153	82,855	285,298
l. Commentaries/columns/reviews (broadcast)	145,553	65,386	80,167
m. Ceremonious speeches/sermons	18,075	12,510	5,565
n. Lectures/seminars	140,901	79,067	61,834
o. Read speech	903,043	351,419	551,624
Total	8,940,098	3,285,631	5,654,644

Table 3.1: CGN database information

3.3.3 JASMIN-CGN

The JASMIN-CGN (*Jongeren, Anderstaligen, Senioren en Machine-Interactie voor het Nederlands*) is an extension of the Spoken Dutch Corpus with speech of elderly people, children and non-natives in the human-machine interaction modality ([Cucchiarini et al. 2008]). It contains 90 hours of recordings (complementary speech for CGN) where 9h 26m are recorded

by native adults above 60 years old in the Netherlands. The main interest on the acquisition of this database is the use of the recordings for adaptation to elderly Dutch speakers.

3.3.4 SmH v1

This database was built in a scenario at Smart Homes [1], with 22 speakers (21 elderly) acting.

Around 1h of continuous sound recordings are provided for each speaker. The recordings are done using multiple microphone channels simultaneously and at different locations of the trial site. The related data aquisition information is described at section B.1.

3.4 Discussion

The efforts of data acquisition were hard. Not only this, in order to proceed with the use of all data (for training, adaptation or test purposes), a data-mining work was required. It was detected the existence of wrong transcriptions and sound recordings, specially when dealing with the large databases. It is a fact, for the author, that the confidence of the data's transcriptions and the quality of recordings is directly related to its usability or not in HMM's training or adaptation campaigns.

[1] Smart Homes is a Dutch partner of the CompanionAble Project, who provided a home designed to be integrated with the best life assistance technologies (http://www.smart-homes.nl/).

4 *Acoustic Modeling*

The main Acoustic Modeling Techniques are described in this section. Acoustic Modeling concerns the creation and the adaptation of the Hidden Markov Models, which in this case represent basic speech units (phones). First, the classic adaptation methods are described in section 4.1. After this, the section 4.2 describes the Multilingual Acoustic Modeling techniques. It is known that this subject commonly has different terminologies documented for each approach (nomenclatures like data sharing, cross-lingual transfer, adaptation, porting, bootstrapping, language independent, language adaptive etc are used and sometimes one specific subject is called by multiple names). This chapter not only explains the multilingual approaches but organizes them in a better understandable and hierarchical tree. It also provides advantages to use the methods.

4.1 Adaptation Methods (MAP *versus* MLLR)

4.1.1 Maximum A Posteriori (MAP)

The MAP adaptation has a capability of achieving performance near to SD (speaker dependent) systems [Goronzy e Kompe 1999]. It uses effective combination of prior knowledge, i.e. the initial model parameters, and ML estimates obtained on the adaptation data (as shown by [Goronzy e Kompe 1999, Zavaliagkos, Schwartz e McDonough 1996]). The MAP adaptation can also deal with foreign accents where often some phonemes differ a lot from the usual pronunciation while other phonemes don't.

The main disadvantage of MAP is the large amount of adaptation data needed before all phonemes can be updated. This adaptation may be hard for those phonemes which do not appear very frequently in the adaptation database.

4.1.2 Maximum Likelihood Linear Regression (MLLR)

The adaptation of HMM based speech recognition systems using Linear Regression was initially introduced by Mokbel [Mokbel 1992].

This method is commonly used for not-supervised adaptation. Regression trees and transformation matrices are used in order to adapt the means and variances of the acoustic models using little adaptation data. The MLLR finds a transformation which maximizes the adaptation data likelihood. The transformations can use less parameters (shared between many phonetic units) and they are, therefore, more robust to recognition errors [Padmanabhan e Picheny 2002].

The basic principle is to calculate one or several transformation matrices from the adaptation data. Clustering several Gaussians into regression classes that share the same tansformation or regression matrix creates the possibility to update even the non observed parameters (i.e. a parameter that is not observed will join the group of an observed one, which has nearer acoustic features, and use the same transformation). The MLLR can quickly adapt the acoustic model to new speakers, environments, channels etc requiring only few adaptation data. One limitation of MLLR is about the foreign accents: Some phonemes will differ a lot from the usual pronunciation and the use of a transformation matrix in the specific regression class of this phomene may not be appropriated for acquiring better results.

4.2 Multilingual Acoustic Modeling

4.2.1 Introduction

Multilingual Acoustic Modeling (MAM) is a large research subject. Nowadays, in most of the Automatic Speech Recognition systems, before recognizing any speech from a specific language, an acoustic model should be built. The way that this acoustic model is built will make a big difference on the speech recognition accuracy, developing time and project costs. Therefore, this section presents some main approaches for this task. The choice of the best approach is directly related to the availability of data, human language expertise and developing time.

4.2.2 Objectives

Several techniques about MAM have been proposed and the results obtained are very interesting for the speech recognition research community. The main idea is to work with a universal

(multi-lingual) phone set to model a target language. When using multiple languages to build the universal set, the acoustic information will be more precise and the recognition of a specific sound in this universe should be better.

It has been proved that even for those who work only with monolingual speech recognition systems, some of the the MAM techniques can provide better recognition results (in recognition accuracy and developing time).

4.2.3 Main Techniques

MAM is studied here to better deal with multi-language association approaches, which can be very useful to speed up the development of an Automatic Speech Recognition system for a new target language. The figure 4.1 shows how MAM can be divided into two main categories which are called Language Independent Modeling and Porting.

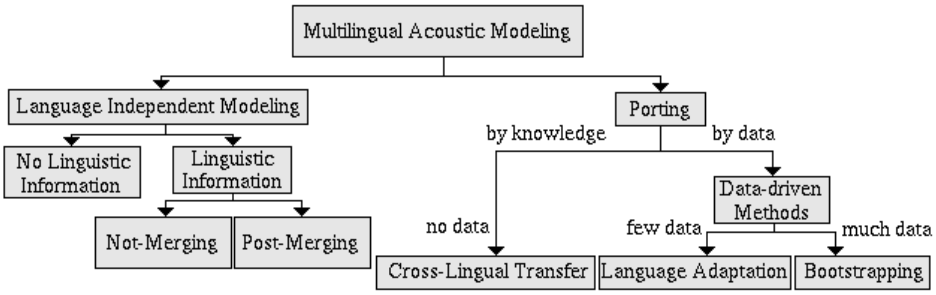

Figure 4.1: Multilingual Acoustic Modeling Approaches

4.2.3.1 Language Independent Modeling

In this category, a reduced model set can be created by a collection of monolingual model sets. The possibility of training data with the sources separated by language and making phoneme model fusions will distinguish three possible ways of using this type of modeling, namely Non Linguistic Information, Not-Merged and Post-Merged.

- No Linguistic Information

  Here, resource sharing is done. All data (or partial data) from the source languages are used together to train the models, with no linguistic information. Everything is transcribed to a standard international phone set like the International Phonetic Alphabet (IPA). Which means that even if the same phone is shared across two different languages it will anyway train the same model. This can be treated as some sort of automatic merging technique, which works just by the fact of ignoring linguistic information.

- Linguistic Information

  Not-Merged

  In this case, resource sharing with linguistic information is used. Separated models exist even for the same phone (because they are labeled by language). The labeled language models stay separated (a larger phone model set is built).

  Post-Merged

  Even if no resource sharing is done, a reduced phone set can be built yet by what we will call here Post-Merging. Separated models that are trained for the same phone in different monolingual phone sets can be reduced to a single model. A model fusion by using some clustering techniques is done by [Garcia, Mengusoglu e Janke 2007] and the result is a reduced model set.

4.2.3.2 Porting

When Porting, one or more source languages will acoustically model a target language. The main idea of porting is initially to associate the target language with seed models and then try to do some adaptation, depending on the training data available. The phonetic cross-lingual association have two different manners of being conducted: Data-driven or Knowledge-Based.

- Knowledge-Based

  Each target language phone is associated with a source language phone (acoustic seed models) using pure human linguistic expertise.

  Cross Lingual Transfer

  Training data is not available. Therefore the seed models cannot be trained or adapted to the target language and human knowledge-based approach is used to associate each phone model.

  Recognizing speech after directly transfer the acoustic models to a target language is called “Cross-language recognition by substitution” by [Fung, Yuen e Kat 1999], where English acoustic models were quickly ported up to Mandarin (and adapted in a “Language Adaptive Modeling” way as described further).

- Data-Driven

  The phones are associated using data from the target language. The target language data is recognized using the source acoustic seed models and the association can be done

automatically by some heuristic like choosing the best time aligned matching phone in the target reference transcriptions.

Language Adaptive Modeling

Training data is available (but few) and will be used for adaptation. This approach uses seed models (obtained from one or more other languages in order to increase the phone coverage) and is better than the flat starts or random model monolingual approaches. Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) are the most prominent techniques used for this adaptation ([Schultz e Waibel 1999]).

Bootstrapping

Many training data is available. Therefore other language(s) seed models will be trained by large target language training data to rebuild a new model. This approach (like Language Adaptation) can be better than flat starts or random models [Schultz e Waibel 1999].

As the choice of languages used in the seed models can greatly influence the obtained results [Gokcen e Gokcen 1997, Schultz e Waibel 1999], the use of a coverage coefficient is important to measure the number of target models covered by the models of the universal set. The calculation of the **sound unit sharing** and the **phonetic coverage coefficient** across languages is described in [Schultz e Katrin 2006].

Some studies [Bub, Kohler e Imperl 1997, Constantinescu e Chollet 1997] indicate a relation between language similarity and cross-language performance. Multilingual systems built with cross-lingual transfer of models have also proved to be better than monolingual ones as shown in [Bub, Kohler e Imperl 1997, Schultz e Waibel 1998].

4.2.4 Conclusions

The MAM techniques are potentially useful to quickly build speech recognition systems, in special for languages which have low resources available. Even if the system is developed to provide monolingual speech recognition, the recognition accuracy is improved by using data from different languages on the acoustic modeling (training seed models) step.

Different techniques exist for every situation regarding available databases, linguistic human knowledge and developing time.

5 *Implementations and Experiments*

5.1 Main Implementations

The main implementations are described in this section. The experiments are conducted with the trained acoustic models (HMMs) and with the language models for the Dutch and the French languages. The configuration of the HMMs are listed in the table below:

	French	Dutch
Extracted features	MFCC_D_A_Z_0	MFCC_Z_E_D_A
Models (monophones)	44	39
States (per model)	5	5
States (for the short pause model)	3	3
States (for the silence model)	5	1
Mixtures (per state)	256	32

Table 5.1: Acoustic model's configuration

For the extracted features, the meaning of such letters is explained below:

MFCC	Mel Cepstral Coefficients (default: extracts 12 cepstral parameters, from C_1 to C_{12}).
_D	Deltas (first derivates of all extracted coefficients)
_A	Acceleration (second derivates of all extracted coefficients)
_Z	Cepstral Mean Normalization
_0	Use C_0 as the energy component
_E	Log energy is appended

Table 5.2: Description of the extracted features

The multiple implementations are performed in order to permit the evaluation of recognition results, including the automatic acoustic adaptation and every step concerning the automatic speech recognition using the Julius and HTK systems. Some of the most important implementations are listed below:

- Modification of the Julius Engine

The Julius engine (version 4.1.2, 4.1.3 and 4.1.4) was modified to deal with MFCC extraction calling HCopy (HTK Tool).

A new module called adinnetmfcc was created inside the Julius server, mixing code lines of input from network (adinnet module for the client-server mode) and MFCC file (mfcfile module, capable to do local speech recognition from the already extracted MFCC coefficients).

Feature extraction and decoding are left to be carried out at a remote computer (server).

Therefore the use of the server-client mode is justified by the need of client efficiency and the potential future use of more powerful processors in the server side in order to avoid delays (specially concerning the time of the MFCC extraction step).

- Cross Validation for different adaptation techniques

  The Cross Validation technique program was implemented to permit the experiments described in section 5.2 [Caon et al. 2010].

- Cross-Lingual Porting

  A cross-lingual porting experiment was done trying to port French phones to Dutch, with the use of an association matrix. However the association was simple and had no aid of Classification and Regression Trees(CART), as conducted by the works of Bayeh ([Bayeh 2009]). The cross-lingual association is useful to build (porting) acoustic models for a language with low resources.

- Automatic Generator of Sentences for Speaker Adaptation

  As the target speaker (elderly CR who lives alone) is always the same in his house, the ASR system can be adapted to his voice during installation. With the aim of creating a set of sentences for speaker adaptation in French, a program was created. The program uses as input a list of sentences, a list of phonemes and a dictionary. Inside the program, it is possible to choose the number of sentences desired in each set (default: 10) and the minimum number of mono-phones of each phoneme contained in the list of phonemes that the sets should have (default: 2, to ensure a balanced coverage). The results is a set of 10 sentences in French, which is read once and in less than 5 minutes (so the CR will not feel tired).

5.1.1 Results

A first recognition test was done with the Clean French Distress Expression database. The grammar is deterministic and is made of the distress expressions. The acoustic models were trained on French broadcast news from the ESTER Corpus([Galliano et al. 2006]). Our ASR system achieves 96% of word accuracy on the clean database.

Thus, we test our system on the Noisy French database. Those recordings are 25 seconds long with a short speech part in the middle. In the background, the same noise is present at several SNR levels. With this Noisy French database our ASR system achieves 97.9% word accuracy.

However, as the size of both Clean and Noisy French database is small and considering that there are few speakers and many repetitions of the sentences, the confidence interval is large. Not only this, the Noisy French Distress Expressions database contains much less utterances than the Clean French Distress Expressions database.

Although the results have been obtained using a restricted vocabulary (related to the French Distress Expressions), with acoustic Hidden Markov Models trained on broadcast news, and with no acoustic adaptation, the recognition accuracy is still high.

5.1.2 Perspectives

These preliminary results on small distress expression databases were very promising and needs to be validated on larger databases. The new built database (mixing the Clean Distress Expressions database with noise recordings) and “La Collégiale v1” (described in chapter 3) will be used for further tests with the French language.

Anyway, another challenge in this work is to deal with several languages, not only French. The usual way is to train acoustic models on a huge corpus for each target language. But in our framework, initially it was not possible because only small corpus were be available for each target language, except for French (and recently for Dutch). Thus, studies in the area of Multilingual Speech Recognition (described at section 4.2) may be a solution for this.

5.2 Adaptation Experiments with Cross Validation

5.2.1 Introduction

The objective of this experiment is to find the best adaptation technique (a common research in the area of automatic speech recognition, e.g. [Wang, Schultz e Waibel 2003]) for the French Acoustic Models (HMM's) trained on ESTER [Galliano et al. 2006] broadcasting news database. The French acoustic models are the adaptation targets.

Different databases are tested and the adaptations are supervised. The use of a K-Fold CV aims to provide more reliable evaluation of the results.

The experiments are conducted in a lower level language unit using forced phoneme alignment by Viterbi. This is known as a good tool for identifying the actual pronunciation contained in the utterances (although the best matching pronunciation must be previously listed in the lexicon, as described by [Young et al. 2006]).

The Maximum Likelihood (ML) re-estimation is used with two different configurations: The first updates transitions, means, variances and weights (ML tmvw). The second configuration doesn't update the transitions (ML mvw).

The Maximum *a posteriori* (MAP) and the Maximum Linear Likelihood Regression (MLLR) adaptation methods are the other tested techniques. It is known that HTK's MAP implementation [Young et al. 2006] does not update transition probabilities while the means (m), the variances (v) and the weights (w) are acceptable options. The explanations about the HTK's implementations are documented in [4]. MLLR is used in a static two-pass adaptation approach, which is described further.

The main features of MAP and MLLR adaptation approaches are given in section 5.2.3.

A K- Fold CV is used to evaluate the experiments. More details about this technique will be introduced later in subsection 5.2.3.1. As it is not only a matter of adaptation, but also to have a good evaluation method, the phone or the word aligned comparison with references are studied in 5.2.4.1.

5.2.2 Databases, Models and Objectives

Three different types of databases are employed in this work: Readings, Interviews and Distress Situations. These databases are used to adapt the HMMs initially trained by the ESTER Corpus.

The ESTER [Galliano et al. 2006] database is recorded on French broadcasting news and around 40 hours were used to train the acoustic models (hidden Markov models composed by monophones containing 5 states and 256 mixture components per state). The language model is based on 3-gram probabilities from large newspaper data ("Le Monde") and the dictionary composed of 65k words.

For the databases used in adaptation, the speakers are non-native French, which permits some analysis on non-native speech adaptation.

The Speaker Dependent Interviews (SDI) are recorded by two non-native speakers (SDIm by one male speaker and SDIf by one female).

The FDE contains French Distress Expressions [1] recorded by 19 native and 2 non-native speakers in distress situations. The databases are summarized in table 5.3.

Database	Utterances	Words	Speakers	Speech Type
SDR	162	1572	1	Read Text
SDIf	103	521	1	Interview
SDIm	103	521	1	Interview
FDE	2646	10080	21	Distress Exp.

Table 5.3: Database information

5.2.3 Experimental protocol

5.2.3.1 Validation Technique

Validation Techniques have two main problems in the pattern recognition research area: Model selection and performance validation. Our aim is to validate the performance (recognition accuracy) of adapted acoustic models by taking into consideration the transcription level of references (detailed in 5.2.4.1).

A K- Fold CV was implemented. The variable K is equal to 20 to make the experiments, which means that every time about 5% of the data is tested while 95% (the other K- 1 parts) are used for adaptation. The K- Fold CV technique is commonly used to give more accurate evaluation results. K has to be chosen accordingly to the database size due to the desired computational time issues and the expected bias for the true error rate.

The figure 5.1 explains how the experiments use a supervised K-Fold CV adaptation, where D is the adaptation database, D(k) is the k-th data subset, M(k) is the adapted acoustic model

[1] Author thanks ESIGETEL (http://www.esigetel.fr/) for providing the French Distress Expressions database.

obtained using all the adaptation data available in D except the k-th part D(k) which is a test database.

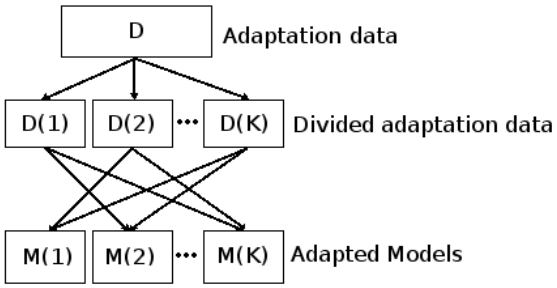

Figure 5.1: Supervised K-Fold CV adaptation

The 20-Fold CV is used to give the mean of recognition accuracy for the not-adapted (original models trained on ESTER corpus) and the adapted acoustic models (with ML, MLLR and MAP options).

5.2.3.2 Two-Pass MLLR

The MLLR adaptation can be supervised (using labeled adaptation data) or not (labeling adaptation data by recognition before adapting the target).

In this work, we used a supervised MLLR, with two-pass static adaptation by means of HERest (HTK tool) [2]. The first one of the two-pass steps builds a global transformation class, while the second pass builds a (multiple) regression transformation class [Young et al. 2006].

5.2.4 Experimental Results

5.2.4.1 Forced Alignment Options

The experiments start by analyzing the best forced alignment options. The alignment type will directly affect the observed results.

To choose the best language unit level to evaluate the adaptation, the SDR database is used to do a ML re-estimation on the ESTER's acoustic models. The other databases could be used to evaluate the forced alignment options too, but SDR was chosen because it is not as large as FDE and not as short as the SDI. Also, a ML re-estimation should give good results with the SDR database, due to the utterance's repetitions and the sufficient data covering all the

[2] HERest is the main HTK's training tool. It performs a single re-estimation of all HMMs, simultaneously. It uses the Forward-Backward algorithm to store statistics of state occupation, means, variances etc.

acoustic model set. For building the phoneme reference transcriptions from word reference transcriptions, the procedure is simple: The first occurrence of possible pronunciation for each word found in the dictionary is chosen. This makes the reference files not always compatible with what will be truly spoken. For example the French *liaisons* are spoken in an random mood (the speaker sometimes makes the *liaisons*, and sometimes it doesn't). This problem could be solved by a manual review of the reference transcriptions, although it can be very hard and expensive for larger databases.

In this work, the experiments employ an adaptation procedure of transcriptions which permits to fix some mispronunciations (if provided in the lexicon, the best pronunciation can be chosen by HTK's Viterbi tool), but the reference file construction (considering only the first pronunciation on the lexicon of each word) is not very flexible. This way, the adaptation is conducted with no mismatch except the evaluation results which may be affected by the variability of the lexicon's pronunciations for a specific word.

At the same time, for evaluating results with a higher language unit level like words, the chance of error is higher in the hypothesis. This is due to the impact of the language model probabilities (n-grams) to make the system fail even if it recognizes the right phonemes before word aligning.

If an utterance contains "there for" (spoken fasten as "therefore", with no short silence interval between the words), the pronunciation will be "DH EH R F AO R", the same for the word "therefore". Then, the hypothesis depends on the language model probabilities for choosing if the recognized output (word aligned) will be "there for" or "therefore".

The results considering word and phoneme forced alignments are presented in figure 5.2. The Phoneme Error Rate(PER) is used for phoneme alignment and the Word Error Rate(WER) is used for the word alignment.

The results confirm a better recognition rate for phoneme aligning (PER is lower than WER). It cannot prove that phone alignment is better than word alignment for evaluation though. The last column shows the error rate mean of all folds together.

The goal is to provide a more reliable information by the use of the K-Fold CV, instead of taking just one random fold to validate. Although the inflexibility of the reference transcriptions at phoneme level (as described before) the experiments results are given in Phone Error Rates (PER) and should provide good observations about adaptation gains. The PER makes us more independent from the language model probabilities. This is a very good aspect as we assume to

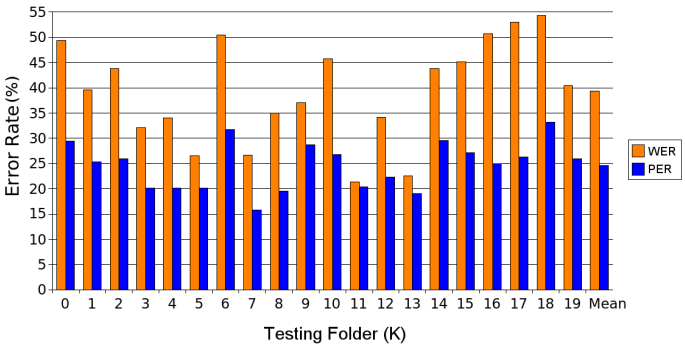

Figure 5.2: SDR's comparison of error rates after ML re-estimation, using 44 models and more than 54k words in the dictionary

analyze only the acoustic model adaptation.

5.2.4.2 Adaptation Experiments

The tests are conducted using the original HMMs trained on the ESTER database (Not-Adapted) and the HMMs adapted with different techniques (ML, MAP and MLLR). The different configurations are listed below and the results illustrated at figure 5.3.

- **ML tmvw** Maximum Likelihood *a priori* re-estimation of transitions, means, variances and weights;
- **MAP mvwp** Maximum *a posteriori* adaptation of means, variances and weights;
- **ML mvw** Maximum Likelihood *a priori* re-estimation of means, variances and weights;
- **MLLR 2-pass** Maximum Likelihood Linear Regression adaptation (as explained in 5.2.3.1).

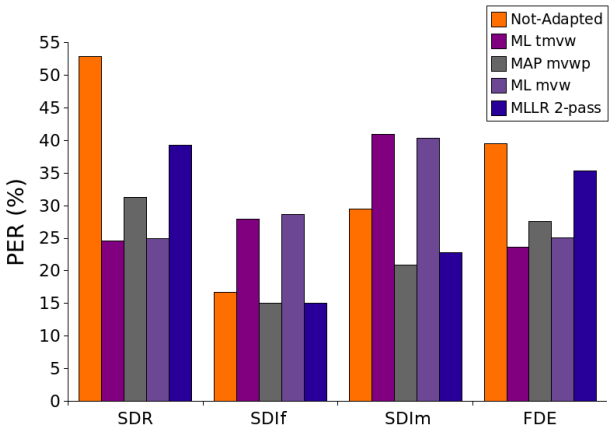

Figure 5.3: Adaptation results for different techniques.

The results are shown by means of the 20 folds (results added and divided by 20, the total fold number) for easier visualization. Due to the high variation of the recognition accuracy results for each part, the mean of the K recognition tests is an information more useful than taking only a random part which explains the use of K-Fold CV.

It can be noticed from figure 5.3 that ML has a better impact in the SDR and FDE databases than MAP or MLLR. This is explained by the fact that the SDR and the FDE have many utterance's repetitions (around 5 for SDR and 21 for FDE). For the SD Interview databases , a better accuracy is observed when doing MAP or MLLR than when doing ML re-estimation.

5.2.5 Conclusion

Experimental results by means of word error rate (WER) and phoneme error rate (PER) indicates that each adaptation method depends differently on the adaptation data, and that the acoustic models performance can be improved by the use of alignments at phoneme-level and K-Fold Cross Validation (which identify the best technique to apply in each data type).

The use of phoneme's alignments is recommended for the evaluation of acoustic models adaptation. It is still better than considering, for example, the comparison of two words with the same pronunciation as being mismatched. The adaptation method should be chosen according to the data available. If the data is sufficient for covering the acoustic space or there are mispronunciations (like in foreign accents), MAP is better. With enough statistical information about the acoustic space and not too much mispronunciations, MLLR is effective even with short adaptation data.

6 System Demonstration

6.1 Introduction

After conducting experiments like adaptation, cross-validation and having done an update in the speech recognition engine (Julius), the next objective is to make a demonstration of the system, testing it with a more realistic situation but still under a controlled scenario.

The first demonstration scenario (described at A.2) was hold at Smart Homes (Eindhoven, Netherlands). Thus, the speech recognition system had to be rebuilt to work with the Dutch language (instead of the French language which was used until this task). The configuration of the used HMMs (acoustic models) has been described in the table 5.1.

The section 6.2 explains what had to be done to achieve the ready integrated system of speech recognition for Dutch language in a short time. Section 6.3 provides the error rates obtained after concluding the first prototype of the Dutch ASR system.

6.2 Requirements

The Dutch demonstration is a first step trying to prove that soon we will be able to recognize speech 24 hours a day, 7 days per week. The speaker targets (care recipients) are Dutch elderly speakers living alone.

In order to achieve a Dutch Demonstration of the speech recognition engine, the usability of speech databases for the Dutch language and the reasonable ways of creating acoustic and language models (to feed up the speech recognition engine) had to be studied. Finally, the Integration with other modules was concluded.

6.2.1 Language Modeling

The language modeling was done using the previewed sentences for the first scenario. Different n-grams and grammars were tested in order to determine which one had the best recognition rate.

6.2.2 Acoustic Modeling

As the main target language is Dutch, the available databases are the Groningen Corpus, CGN and JASMIN. There is also a shorter database specially recorded for the CompanionAble project in Smart Homes[1]. This database (that we will call 'SmH') contains 21 target speakers to whom the acoustic models were adapted to improve and test the speech recognition accuracy.

For all the chosen databases, the transcription's accuracy and other data specific information like speaker's age, speech type (e.g. dialogue, read text etc.) and sound quality were taken into consideration to determine how to build the Dutch acoustic models.

The databases identified as most useful for building a Dutch acoustic model are the CGN and its extension, called JASMIN. The Groningen Corpus has some inconsistencies of transcriptions and doesn't use the same standards, meaning that it may take more time to be prepared.

Therefore, we have 3 databases to manage acoustic modeling in Dutch:

- Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN)
- JASMIN-CGN - Jongeren, Anderstaligen, Senioren en Machine Interactie voor hetNederlands
- SmH . Smart Home's recordings.

The CGN data can be filtered by conversation type and age (elderly) to build the first acoustic models. That means only the database files which contain specific group of speakers are selected as training data. After this, Jasmin filtered speech data is used to adapt the first models obtained from CGN. Finally, the acoustic models are adapted to the target speaker (care recipient) contained in the SmH database. Adaptation is conducted using MLLR technique. MAP technique was also used for experimental purposes of comparison.

The initial HMM adaptation steps are illustrated in the figure 6.1:

[1] http://www.smart-homes.nl/

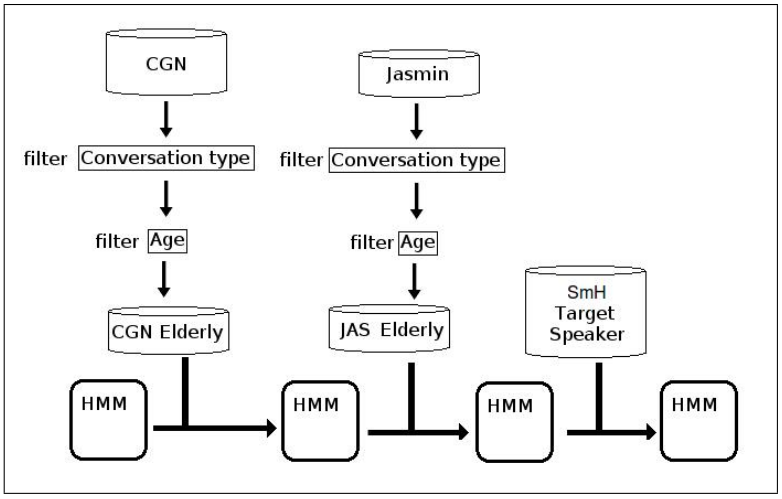

Figure 6.1: Schema for acoustic models's learning and adaptation.

The schema shown in 6.1 was identified as being the recommended way to create an acoustic model for an elderly Dutch speaker. However, in the project's schedule, an younger Dutch speaker has been selected to do the Dutch demonstration. Thus, the first speaker adapted model was built directly by doing adaptation to a CGN acoustic model trained with read-text (a specific component of CGN Corpus) to the demonstrator's voice (recorded in the SmH database).

6.2.3 Integration with support systems

The ASR system is an interface of communication between the care recipient, the companion robot and the house. It works together with the other modules like the dialog manager.

Initially, the capability to distinguish speech from the other sounds came only from the Julius Engine's VAD (voice activity detection), although it is experimental. A second option was chosen and now the audio signal is acquired from a second system (a sound module), developed by ESIGETEL, which is able to classify sound types and provide only speech signals to the speech module.

It was decided that all the project's system modules should send their outputs and request inputs from a SABB (System Awareness Black Board) using SOAP connection.

Therefore, work was conducted to integrate the speech module to the blackboard, in order to allow a further dialog manager to deal with the recognized sentences.

6.3 Results

The first phase of integration was concluded in time and the first results obtained.

For this initial version, the young speaker was chosen to demonstrate the system, therefore the acoustic models (trained with the CGN read text component) were directly adapted to its voice (recorded on the SmH database). After many system's configurations, choosing if n-grams or grammars should be used and testing if adaptation by MLLR or MAP would give better results, the best result was obtained with MLLR and n-grams (with fillers to avoid the sentences which were out of the scenario's scope).

The table 6.1 shows the one of best results obtained with the ASR system, after speaker acoustic adaptation using the MLLR approach and n-grams (with fillers) as language model. The speaker repeated 10 times each one of the sentences. The results appear as CCR (Completely correctly recognized) and SCR (Semantically correctly recognized). The semantical rate considers "Hello" and "Hi" have the same meaning, so it is not an error if they are switched during recognition.

It is important to note that the VAD developed by ESIGETEL was integrated only after the rates (provided in table 6.1) were obtained. Moreover, the rates analyzed only the 1-best output of the ASR module. This number can be easily increased at run-time. The dialog manager module has the potential to work with more hypothesis and provide more accurate answers to the care recipient's needs.

6.4 Conclusion

A reasonable way was applied to develop a speech recognition system for a demonstration, running in Dutch in very short time. In order to deal with most of the occurred errors, it is expected from the dialog manager to have the artificial intelligence to know how to judge when the secondary hypothesis of recognition can activate voice commands or awareness situations.

The SmH database can be used as target to test all Acoustic Models which are built using different techniques and combinations of databases.

The 21 elderly (and 1 young) speakers from SmH have recorded not only scenario utterances, but also a set of speaker adaptation sentences. That is very useful for the research, providing much confidence about results obtained in a real situation.

The Dutch language model was built first in a simple version, following the demonstration's

Table 6.1: Results after MLLR speaker adaptation, using n-grams and fillers.

Utterance	CCR(%)	SCR(%)
hellep	100	100
help me	50	90
kom naar de keuken	100	100
kom eens naar de keuken	100	100
wil je naar de keuken komen	60	60
wil je m'n bril aangeven	80	90
geef m'n bril eens aan	100	100
wil je m'n bril brengen	90	100
breng me mijn bril	100	100
tot ziens	90	90
houdoe	100	100
ja graag	70	100
graag bedankt	70	100
ne dank je dat is alles	50	90
nee dat is alles	90	100
nee dat was het	100	100
nee	100	100
ja graag	50	80
graag	70	100
ja	100	100
kan je mij met hem doorverbinden	100	100
wil je me met hem doorverbinden	100	100
ik wil graag mijn geheugen trainen kan je het programma opstarten	70	80
hector ik wil graag mijn geheugen trainen kan je het programma opstarten	100	100
wolly ik wil graag mijn geheugen trainen kan je het programma opstarten	70	70
hector wil je de geheugentraining opstarten	70	70
dat ben ik vergeten bedankt voor de herinnering ik zal ze nu innemen	90	100
helemaal vergeten dank je dat je het me helpt onthouden ik zal ze nu innemen	100	100
hector ik ga lunchen met vrienden	100	100
hector ik ga uit eten met vrienden	100	100
hector ik ga lunchen met kennisen	100	100
hector ik ga lunchen met buren	100	100
wolly ik ga uit eten	100	100
wolly ik ga uit eten met vrienden	100	100
wolly ik ga uit eten met kennisen	100	100
ja graag	40	80
Average percentage	86.39%	94.44%
Lowest percentage	40%	60%
Highest percentage	100%	100%

scenario, and later can be expanded accordingly to the recognition accuracy obtained.

7 Conclusions and Future Work

There are multiple ways to increase speech recognition rates, and that was mainly what this research aimed to do. The experimental results in French language showed gains around 15% in phoneme accuracy (mean of 20 folds using the k-fold cross validation technique) with the classical methods of adaptation (MLLR and MAP) and re-estimation (ML) of the model parameters using the French Distress Expressions database. The Dutch language became a priority during the works, and the results of speech recognition after acoustic adaptation (from the models learned with the CGN component-o read-text recordings) to a specific speaker (and the creation of language models for a specific scenario to demonstrate the system) showed 86.39 % accuracy rate of sentence for the Dutch acoustic models. The same data showed 94.44 % semantical accuracy rate of sentence. It is important to note that the rates obtained didn't make use of the independent Voice Activity Detection (VAD) system (integrated later).

The objectives were not only a matter of increasing accuracy, but also to have confident recognition rates, therefore the K-Fold cross validation program was implemented.

As the scientific contribution, the Julius Engine was modified in its three last versions in order to fix the feature extraction from live microphone input and sound files. This version is now integrated to an independent Voice Activity Detection (VAD) system to filter non-speech inputs.

Our system was capable to recognize speech in French and Dutch languages. The desired language was set in run-time by the choice of the acoustic model and language model. The recognized sentences were directly registered in a blackboard with date-time information and scores to allow the dialog manager to deal with this input.

The perspectives, in terms of research, points to the Multilingual Acoustic Modeling studies as a potential source of improvements concerning alternative methods, to recognize multiple languages instead of the standard methods, which require large sound databases and more developing time.

Finally, when someone listens, normally it makes a fusion between its "body sensors",

specially the visual information, to complement the comprehension of speech. An example of this is the labial recognition of speech made by vision. One can presume other one is talking if it is opening its mounth. This information can be investigated, seeking the configuration of a more precise voice activity detector.

8 Acknowledgment

It was a great pleasure for me to do this Master's Internship at Télécom Paristech, also with the cooperation of Télécom SudParis and ESIGETEL.

First of all, thank's to God, for everything.

I would like to thank my family, for my education and the love.

I would like to thank my main advisor in France, the professor Gérard Chollet, who gave me very good instructions when needed and who let me free to choose the tasks which I should deal with priority.

I would like to thank my co-advisors in France, the professors Jérôme Boudy (Télécom SudParis) and Dan Istrate (ESIGETEL), for the very fruitful discussions and great assistance concerning the integration of our system modules and work synchronization.

I would like to thank the professor Rodrigo Varejão Andreão, co-advisor of my Mastership in Informatics in the UFES (Universidade Federal do Espírito Santo, in Brazil) for being so helpful since the first researches about speech recognition with the HTK system in the year of 2008.

I would like to thank the professor Thomas Walter Rauber, the main advisor of my Mastership in Informatics in the UFES, for the comprehension of my best intentions regarding the specialization in the area of Informatics while researching speech signal processing.

I would like to thank those who also worked more near me at Télécom Paristech during my research, specially the professors Joseph Razik and Asmaa Amehraye who provided me great assistance from the beginning to the end of my works.

I would like to thank the professors Rania Bayeh, Leila Zouari, Chafic Mokbel and everyone who somehow contributed to my research providing assistance, although the distance between the continents and the very few times we could see each other personally.

I would like to thank all the partners of CompanionAble Project for being part of this great

experience.

I would like to thank the professor Flávio Miguel Varejão (UFES) and professor Carlos Alberto Ynoguti (INATEL) for having accepted to evaluate this work.

The CompanionAble project is co-funded by the European Commission under the 7th Framework Programme (Grant Agreement Number 216487). Part of this work has also been conducted by the Short Term Scientific Mission, COST Action 2102, n. 18.

Glossary

ASR Automatic Speech recognition.

CR Care recipient.

FP7 Seventh Framework Programme.

HMM Hidden Markov Model.

HTK Hidden Markov Model Toolkit.

LVCSR Large vocabulary continuous speech recognition.

MAM Multilingual Acoustic Modeling.

MAP Maximum *a Posteriori*.

MFCC Mel-Frequency Cepstral Coefficients.

ML Maximum Likelihood.

MLLR Maximum Likelihood Linear Regression.

PER Phoneme Error Rate.

WER Word Error Rate.

Bibliography

[Baldinger et al. 2004]BALDINGER, J.-L. et al. Tele-surveillance system for patient at home: The mediville system. In: MIESENBERGER, K. et al. (Ed.). *Computers Helping People with Special Needs*. Springer Berlin / Heidelberg, 2004, (Lecture Notes in Computer Science, v. 3118). p. 623–623. 10.1007/978-3-540-27817-7_59. Disponível em: <http://dx.doi.org/10.1007/978-3-540-27817-7_59>.

[Bayeh 2009]BAYEH, R. *Reconnaissance de la Parole Multilingue: Adaptation de Modeles Acoustiques Multilingues vers une langue cible*. Tese (Doutorado) — TELECOM Paristech, 2009.

[Benesty, Sondhi e Huang 2008]BENESTY, J.; SONDHI, M.; HUANG, Y. *Springer Handbook of Speech Processing*. [S.l.]: Springer, 2008.

[Bub, Kohler e Imperl 1997]BUB, U.; KOHLER, J.; IMPERL, B. In-service adaptation of multilingual hidden-markov-models. In: *ICASSP*. [S.l.: s.n.], 1997. v. 2, p. 1451–1454.

[Caon et al. 2010]CAON, D. et al. Experiments on acoustic model supervised adaptation and evaluation by k-fold cross validation technique. In: ISIVC. *5th International Symposium on I/V Communications and Mobile Networks*. Rabat, Morocco: IEEE, 2010. To be published.

[Clement, Tennant e Muwanga 2010]CLEMENT, N.; TENNANT, C.; MUWANGA, C. Polytrauma in the elderly: predictors of the cause and time of death. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, v. 18, n. 1, p. 26, 2010. ISSN 1757-7241. Disponível em: <http://www.sjtrem.com/content/18/1/26>.

[Constantinescu e Chollet 1997]CONSTANTINESCU, A.; CHOLLET, G. On cross-language experiments and data-driven units for alisp (automatic language independent speech processing). In: *IEEE Workshop on Automatic Speech Recognition and Understanding*. Santa Barbara, CA, USA: [s.n.], 1997. p. 606–613.

[Cucchiarini et al. 2008]CUCCHIARINI, C. et al. Recording speech of children, non-natives and elderly people for hlt applications: the jasmin-cgn corpus. In: CHAIR), N. C. C. et al. (Ed.). *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008. ISBN 2-9517408-4-0. Http://www.lrec-conf.org/proceedings/lrec2008/.

[Davis e Mermelstein 1980]DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, v. 28, n. 4, p. 357–366, 1980. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1163420>.

[Fung, Yuen e Kat 1999]FUNG, P.; YUEN, M.; KAT, L. Map-based cross-language adaptation augmented by linguistic knowledge: from english to chinese. In: EUROSPEECH. Budapest, Hungaria, 1999. Disponível em: <citeseer.ist.psu.edu/fung99mapbased.html>.

[Galliano et al. 2006]GALLIANO, S. et al. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In: LREC. [S.l.], 2006. p. 315–320.

[Garcia, Mengusoglu e Janke 2007]GARCIA, E. G.; MENGUSOGLU, E.; JANKE, E. Multilingual acoustic models for speech recognition in low-resource devices. In: *icassp*. [S.l.: s.n.], 2007. v. 4, p. 981–984.

[Gersho e Gray 1992]GERSHO, A.; GRAY, R. M. *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

[Gokcen e Gokcen 1997]GOKCEN, S.; GOKCEN, J. M. A multilingual phoneme and model set: toward a universal base for automatic speech recognition. In: *IEEE Workshop on Automatic Speech Recognition and Understanding*. Santa Barbara, CA, USA: [s.n.], 1997. p. 599–605.

[Goronzy e Kompe 1999]GORONZY, S.; KOMPE, R. In: SONY RESEARCH FORUM 99. *A Combined MAP + MLLR Approach for Speaker Adaptation*. [S.l.], 1999. v. 9, p. 9–14.

[Lee e Kawahara 2009]LEE, A.; KAWAHARA, T. *Recent Development of Open-Source Speech Recognition Engine Julius*. 2009.

[Lee, Kawahara e Shikano 2001]LEE, A.; KAWAHARA, T.; SHIKANO, K. In: EUROSPEECH. *Julius - an open source real-time large vocabulary recognition engine*. [S.l.], 2001. p. 1691–1694.

[Mokbel 1992]MOKBEL, C. *Reconnaissance de la parole dans le bruit: bruitage/débruitage*. Tese (Doutorado) — Ecole nationale supérieure des télécommunications, 1992.

[Padmanabhan e Picheny 2002]PADMANABHAN, M.; PICHENY, M. *Large-Vocabulary Speech Recognition Algorithms*. 2002.

[Rabiner 1989]RABINER, L. A tutorial on hidden markov models and selected applications in speech recognition. In: . [S.l.]: IEEE, 1989.

[Rabiner e Juang 1993]RABINER, L.; JUANG, B. Fundamentals of speech recognition. In: . Englewood Cliffs: Prentice Hall, 1993.

[Razik 2007]RAZIK, J. *Local and frame-synchronous confidence measures for automatic speech recognition*. Tese (Doutorado) — Nancy Université I, 2007.

[Russell e Norvig 2002]RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach (2nd Edition)*. 2. ed. Prentice Hall, 2002. Hardcover. ISBN 0137903952. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0137903952>.

[Schultz e Katrin 2006]SCHULTZ, T.; KATRIN, K. *Multilingual Speech Processing*. [S.l.]: Elsevier, 2006.

[Schultz e Waibel 1998]SCHULTZ, T.; WAIBEL, A. Multilingual and crosslingual speech recognition. In: DARPA WORKSHOP ON BROADCAST NEWS TRANSCRIPTION AND UNDERSTANDING. [S.l.], 1998.

[Schultz e Waibel 1999]SCHULTZ, T.; WAIBEL, A. Language adaptive lvcsr through polyphone decision tree specialization. In: *In Proc. the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST)*. [S.l.: s.n.], 1999. p. 97–102.

[Walker et al. 2004]WALKER, W. et al. *Sphinx-4: A flexible open source framework for speech recognition*. [S.l.], 2004. Disponível em: <http://cmusphinx.sourceforge.net/sphinx4>.

[Wang, Schultz e Waibel 2003]WANG, Z.; SCHULTZ, T.; WAIBEL, A. Comparison of acoustic model adaptation techniques on non-native speech. In: *in ICASSP 2003. IEEE*. [S.l.]: IEEE, 2003. p. 540–543.

[Young et al. 2006]YOUNG, S. J. et al. *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[Zavaliagkos, Schwartz e McDonough 1996]ZAVALIAGKOS, G.; SCHWARTZ, R.; MCDONOUGH, J. Maximum a posteriori adaptation for large scale hmm recognizers. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, IEEE Computer Society, Los Alamitos, CA, USA, v. 2, p. 725–728, 1996.

APPENDIX A – CompanionAble 1st Prototype Scenario

A.1 Morning : Robot Only Scenario

Jasmine comes home from the supermarket. She has been shopping for her breakfast. As she comes through the door, Robby, her CompanionAble robot, greets her: “Hello Jasmine, welcome back, shall I keep your keys?” Jasmine is happy to put them into Robby’s box as she can be sure that they are safe and Robby always remembers where they are. For her this has become a bit more difficult for the last couple of years now, as she has been diagnosed with Mild Cognitive Impairment. That was when her daughter, Melanie, who lives in a nearby city, decided to invest into the CompanionAble robot, that will help Jasmine to remember important dates and appointments, as well as giving her company and supporting social inclusion with its in-built video-conferencing system.

After Jasmine has removed her coat she is going to the kitchen to prepare her breakfast, scrambled eggs, bacon and toast. While having breakfast Jasmine always reads the morning newspaper, but now she cannot find her glasses. She calls Robby “come here to the kitchen”. Robby appears and waits near the entrance of the kitchen. Jasmine asks Robby for her glasses: “Can you bring me my glasses?” Robby moves towards Jasmine so that she can take out the glasses and moves back to the entrance of the kitchen. After a few seconds he asks whether Jasmine still needs her. Since Jasmine only wants to read her newspaper, Robby retreats to his resting station to recharge its batteries.

While still having her breakfast, her daughter is calling on Robby’s videophone. He looks for Jasmine and finds her in the kitchen, tells her that Melanie is calling and asking whether he should connect the call. Jasmine wants to talk to her daughter so Robby positions himself in front of Jasmine and adjusts the display so that Jasmine can see it optimally. During the conference call Melanie takes control of the robot to drive around the kitchen to see whether everything is ok there. Jasmine does not mind this as her daughter has always been very caring

about her and her needs. After the video call has finished Robby reminds Jasmine of the upcoming appointment with her therapist, Dr. Harper, which is about to start in 3 minutes. Jasmine asks Robby to connect to Dr. Harper.

Later that day Jasmine is watching TV, but gets a bit bored as her favourite show is not starting until 12 noon. To entertain herself she wants to do some cognitive training and asks Robby to start the programme. The display shows the selection menu from which Jasmine is selecting the game.

Just before lunch Robby is asking Jasmine whether she has taken her medicine. Again due to her condition she has to take medicine regularly but often forgets it which is why she appreciates the gentle reminders from Robby.

As usual she goes out for lunch to meet up with some friends. When leaving the house, Robby follows her to the door and asks Jasmine whether she would need her glasses, keys, and mobile phone. Jasmine takes them out of Robby's box and says "Good bye, see you later" to Robby. Robby retreats to his resting station to await the return of Jasmine.

A.2 Afternoon - Smart Home Only Scenario

After lunch Jasmine comes home to her CompanionAble Smart Home. The home supports her in her daily activities and is adapted to her personal preferences. When she enters the home, the lights are automatically switched on as she walks through the hallway and living room to the kitchen.

As the intelligent pill box installed in her home still has the medicine in it that was supposed to be taken just after lunch, the smart home displays the message "Please take your lunch medicine, Jasmine". Alternatively a sound can be played to remind Jasmine about her medicine.

When in the kitchen, Jasmine suddenly feels dizzy. She drops her glass of water which breaks on the floor, and she shouts help. The Smart Home, having recognised this critical situation, initiates the avatar which directly asks Jasmine whether she is alright. When Jasmine does not answer, the Smart Home sends a message to Melanie her daughter's mobile phone, who straight away calls Jasmine on her phone. When Jasmine picks up the phone and says that she is fine, Melanie is relieved, but still wants to visit her later that evening to check that everything is OK.

After Jasmine has prepared her afternoon tea she goes to the living room to watch some TV. However, as she has forgotten to turn off the tap, a message is already displayed on the Screen to remind her of that. She goes to turn it off, thinking that the system has just saved her some money as she has a water meter. Last time she forgot to turn off the tap the water was running for 6 hours.

Later on Jasmine wants to call her old friend Francesco whom she has known since childhood. He also has a video conferencing system installed, so Jasmine selects the videoconferencing button on the tablet PC next to her and starts the conversation with Francesco, which is displayed on the TV. Francesco is telling her about his progress in his cognitive training programme, teasing Jasmine that he is much better than her. Of course Jasmine wants to catch up with him, so after the conversation she starts the cognitive training programme using the TV.

Since Melanie has her own key to Jasmine's apartment she does not need to ring but simply enters the flat. Again the lights are automatically switched on, which also notifies Jasmine that Melanie has entered the flat. Together they have dinner and decide that Jasmine could stay over at Melanie's place for a couple of days as Melanie's husband Jack is on a business trip and she is alone with her two children. When leaving the house, Jasmine switches the system off completely to save energy, something she can do at any point in time if something goes wrong.

APPENDIX B – Specification of Scenarios For Speech Data Acquisition

The diversity in the data can be increased by means of:

- •different activities
- •different speech expressions
- •moving the microphones to different positions,
- •changing environment conditions (open/close blinds and doors, adding noises e.g. opening newspaper while talking etc),
- •changing position of the speech source.

It is important to make attention while recording, e.g. any hesitation has to be transcribed later.

For the first phase hesitations are not desired.

The database has to be a mix of clean and noisy utterances, that means that while talking the Care Recipient(CR) should be doing something like changing the newspaper's pages sometimes, but not every time, so we can compare different situations.

B.1 Dutch Speech Recordings - Smart Homes

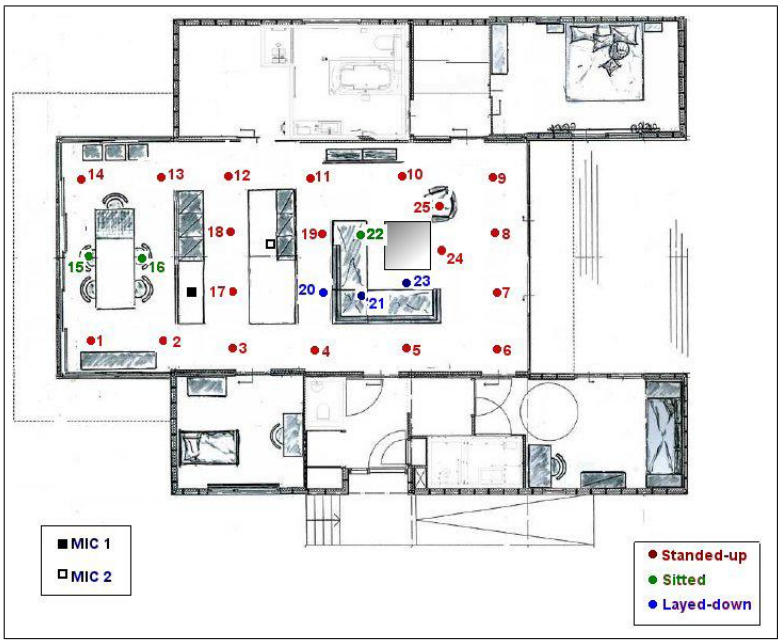

Figure B.1: The floor plan of the smart house with the respective recording spots

The first speech data acquisition concerning Dutch data in a more realistic (confident) situation of scenario was done following the information described in this section.

Initial situation:

- a newspaper has to be put at the guest room
- fruits or cookies at kitchen desk
- coffee ready for preparation, cup in the cupboard.
- some ready (clean) dishes in the dishwasher
- some deco at the desk in dining room

It should be noted that it is not desired to remain in a static situation anywhere, so e.g. when "taking a nap", it is sufficient to just lay down and get up after a few seconds.

For each individual the different scenarios are executed with the defined items:

- **Spot:** Number in the map (1 to 25).
- **Position:** UP standed-up / SIT sitted / DOWN layed-down.

•**Stress:** From 0(weak voice) to 10(disaster shouting).

The tables B.1 and B.2 show examples of scenes prepared for Dutch data acquisition at Smart Homes (there is a total of five different situations). The speech sentences were translated from English to Dutch.

Table B.1: SmH scene 1 - Entering home.

Speech	Action	Place	Position	Stress
Hello...	Say and walk to 5	4	UP	5
Turn on the light!	Look up and say	5	UP	5
I feel cold!	Open the blinds	6	UP	6
More heating!	Face microphone 2	7	UP	6
Thanks...	Say and walk to 9	8	UP	4
Have you seen my glasses?	Say and wait 1 second	9	UP	5
I can't read without it...	Say and walk to 10	9	UP	4
I am not well!	Say and take newspaper	10	UP	6
I feel ill!	Walking to 12	11	UP	7
I am in a bad way!	Walking to 13	12	UP	8
Ouch!	hit the furniture, walk to 14	13	UP	8
It hurts!	Stop and walk to 15	14	UP	8
A nurse, please!	Sit down and call a nurse	15	SIT	7
A nurse!	Continue calling a nurse	15	SIT	7
Quickly!	Say and stand up	15	SIT/UP	7
Bring me a mobile	Say and go to 16	1	SIT	6
Stop listening	Say it while calling	16	SIT	5
Hello doctor, please come...	Say it to mobile	16	SIT	5
I feel dizzy!	Say it to mobile	16	SIT	5
Come quickly!	Say it to mobile	16	SIT	5
OK...	Say it to the mobile and finish call	16	SIT	5
Start listening!	Say it and stand up, go to 2	16	UP	5
Bring me my glasses!	Say and stop	16	UP	5

Table B.2: SmH scene 2 - Having a coffee

Speech	Action	Place	Position	Stress
I'm thirsty...	Say and go to 17	2	UP	4
	Prepare coffe	17	UP	5
Oh!	Hot water in hands (close it)	18	UP	6
It is okay...	Say and wait 1 second	18	UP	5
I'm fine!	Say and wait 1 second	18	UP	5
Don't worry about me.	say and go to 17, 3, 4, 20, 19, 22	18	UP	4
videoconference!	Sit on the couch looking at TV	22	SIT	5
Stop listening!		22	SIT	5
Start listening!		22	SIT	5
end videoconference!		22	SIT	5
I feel so tired.	Lay down on the couch and say	21	DOWN	4
Lower the shutters!	Say and wait 1 second	21	DOWN	5
Pick up the phone!	Wake up and sit, the telephone rings	21	SIT	5
Thanks my friend.	Say, stand up, go to 23	21	SIT	5
Hello...	Talk to telephone, go to 24, 25	23	UP	5
Stop listening!		25	SIT	5
Start listening!		25	SIT	5
I have a headache.		25	SIT	6
Please call a doctor!		25	SIT	6
Stop!		25	SIT	6
I'm okay...		25	SIT	5

B.2 French Speech Recordings - Broca Hospital

The figure B.2 shows the site for the second data acquisition ("La Collégiale" database).

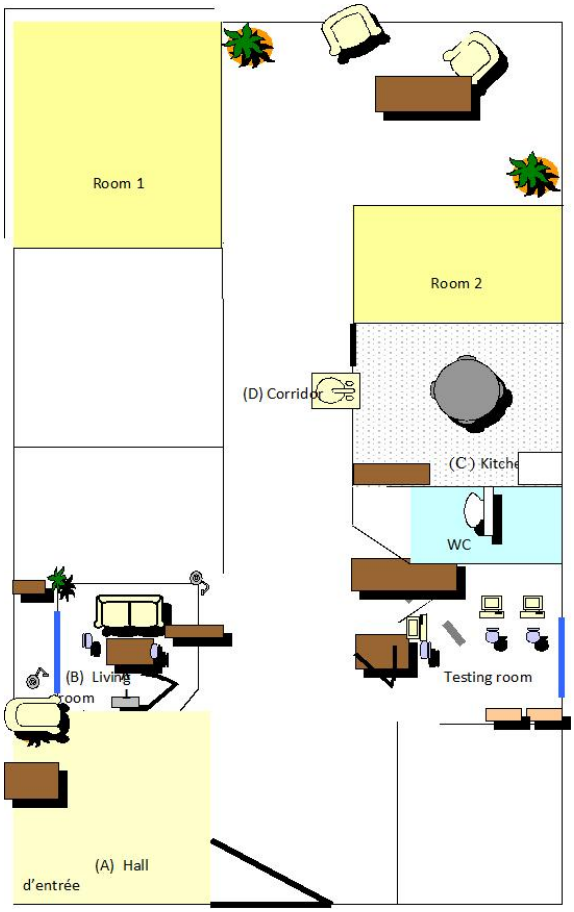

Figure B.2: The floor plan of the Broca Hospital with the respective recording spots