# SPEAKER ADAPTATION TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION

**Supervised Project**

## Bibliography Report

Supervised by

Tugtekin Turan

Written by

Frederic Assmus

Pierre Goncalves

M1 NLP, IDMC Nancy
2019-2020

# Contents

# Chapter 1

# Introduction

## 1.1 Problem Definitions

In statistical speech recognition, there are usually mismatches between the conditions under which the model was trained and those of the input. Mismatches may occur because of differences between speakers, environmental noise, or differences in channels. They should be compensated to obtain sufficient recognition performance. Therefore, acoustic model adaptation is the process of modifying the parameters of an acoustic model used for speech recognition to fit the actual acoustic characteristics by using a limited amount of utterances from the target users. Speech recognition techniques using hidden Markov models (HMMs) have become significantly popular since the late 1980s. In particular, these algorithms often employ continuous density HMMs using triphones as recognition units and a Gaussian mixture distribution as the output distribution. In other words, Gaussian mixture models (GMM) represent the relationship between HMM states and the acoustic input. Over the last few years, advances in both machine learning algorithms and computer hardware have led to more efficient methods for training deep neural networks (DNNs). It has presented in many papers that DNNs can outperform GMMs at acoustic modeling for speech recognition on a variety of datasets including large datasets with large vocabularies. For GMM models, speaker adaptation has proven to be effective in mitigating the effects of this mismatch. In general, it modifies speaker-independent (SI) models towards particular testing speakers or transforms the features of testing speakers towards the SI models. Although displaying superior generalization ability than GMMs, DNN models still suffer from the mismatch between acoustic models and testing speakers. As is the case with GMMs, DNN models experience a degradation of accuracy when ported from training speakers to unseen testing speakers. The use of utterances from many speakers for training enables these models to represent not only phonetic features but also speaker features. Although this ability has made SI systems practical, the systems still do not perform as well as speaker-dependent systems in which the parameters are estimated from a sufficient amount of utterances from one target user. This means that speaker adaptation techniques are very important for any recognition system. In this project, we mainly deal with speaker adaptation to optimize the performance by transforming SI models towards particular speakers or modifying the target features to match a pre-trained SI model based on a relatively small amount of adaptation data from the target speakers. We will analyze different techniques and make a comparison using state-of-the-art adaptation systems.

## 1.2 Report Organization

This report aims to introduce the readings that helped us understand the state-of-the-art around our subject and the subject itself. We will organize this report as follows : First, we will introduce the basics of speech signal processing, which is a general domain that we had to encounter to catch how sound is produced by humans and how it can be represented and analyzed. Then, we will introduce automatic speech recognition, which is the central research domain of our subject. We will discuss what is automatic speech recognition, and how it works with its different components. To go more deeply into our subject, we will finally dive into speaker adaptation. We will explain what it is, what are the issues and what are the methods. To finish, as an extension of the speaker adaptation techniques part, we will explain what are our plans for the next step of the supervised project, which is the realization.

# Chapter 2

# Speech Signal Processing

## 2.1 What is Speech Processing

As we know, speech is the main medium of communication between humans. Speech is sound, and if we want to apply new technologies to speech to make progress in the domains around it, we have to process it. There's two ways of characterizing speech. The first one is the content, which is also called the information and is the message in the speech, the actual meaning of it. The second way of characterizing speech is the signal, which is the actual waveform that carries all the information. One of many problem areas of speech processing is the speech coding (also called speech compression) : to process the speech, one of the most important step is to transform the signal into a representation that will have properties that help us storing and process it in simple ways. To resume, the goal of speech coding is to represent the speech by a minimum number of bits, while keeping the quality of the information.

## 2.2 How Speech is Produced by Humans

Producing speech consists of pushing air from the lungs then making it vibrate in the vocal folds then through the vocal tract, which consists of the laryngeal cavity, the pharynx, the oral cavity, and the nasal cavity. The variation of sounds is obtained given the type of vibration of the vocal folds and the cavity chosen for outputting the sound. A scheme of the general shape of the vocal tract is displayed on figure 2.1.

More specifically, we know what are the physical parameters that create one or another sound. For example, vowels are produced with the vocal folds closed. It's more complicated for consonants, which are produced differently depending on their type : voiced, unvoiced, or semi-voiced. Voiced consonants (like /m/ or /n/) are produced by closing the vocal folds. Unvoiced consonants (like /f/ or /s/) without vocal folds vibration : they result in constriction of the vocal tract, with help of the tongue. Semi-voiced consonants (like /z/ or /v/) are produced like unvoiced consonants, but with closed vocal folds.

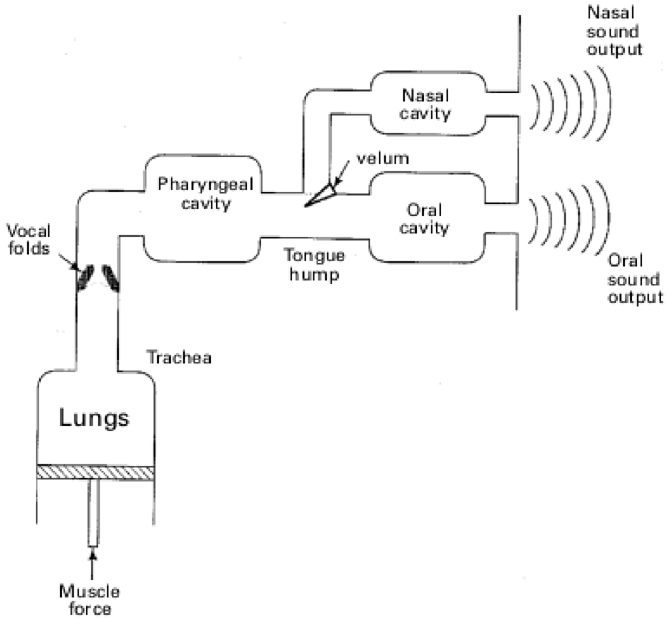

**Figure 2.1:** Vocal tract representation [1]

## 2.3 Speech Signal Representation

We know that speech is represented as a waveform. For example, the waveform of the word "test" can be seen in Figure 2.2 [1].

On this waveform, the horizontal axis is time and the vertical axis is the amplitude. More than just being the display of the signal, this waveform can give us some information about the speech related to it. For example, we already see that there's like a pause around 0.3 s, which corresponds to the pause made when pronouncing slowly the word "test". As the word has been well-pronounced, we can separate the speech signal into 4 parts, each corresponding to one sound (Figure 2.3) [1].

But how did we recognize each sound ? We can see the difference between the only vowel (/e/) and the consonants. For the vowel, the sound is periodic, and that's how we recognize it. We see that consonants have a more « noisy » waveform. To distinguish between /s/ and /t/, it's simple in our case : the /t/'s have the same shape, while the /s/ is different. Here, the main difference between /s/ and /t/ is the explosive character of the phoneme /t/, which can easily be compared to the monotonous character of the phoneme /s/.

Here we've seen that we can deduce the information about the sound by inspecting waveform. We can point out differences between vowels and consonants, and between the explosive or monotonous character they have. The information that we deduce is really basic and is not very precise. This is in part because we look at the waveform in the time domain. To get more information, it's also necessary to look at waveforms in another domain : the frequency domain.

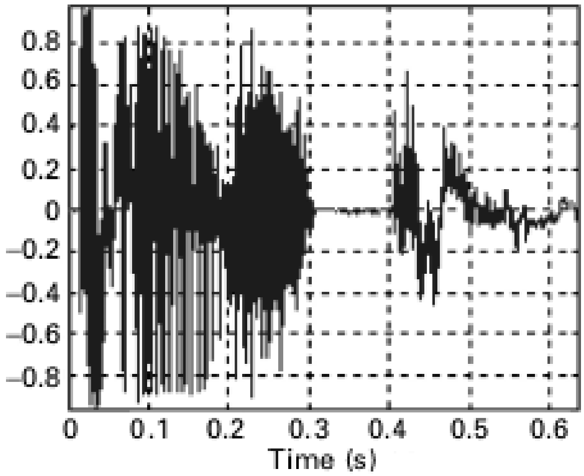

**Figure 2.2:** Speech signal representation for the word "test".

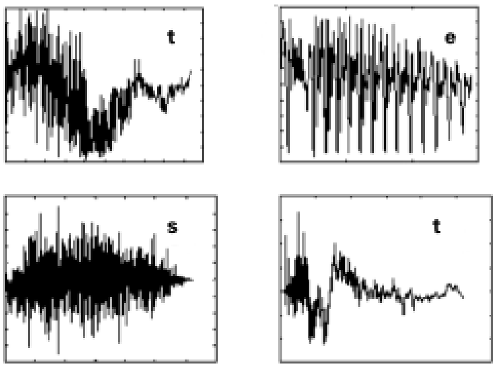

**Figure 2.3:** Speech signal of the 4 phonemes composing the utterance of the word "test"

To switch from time domain to frequency domain, we use a method called Fourier transform, which we will discuss later.

## 2.4 Sources of Variability of Speech Sounds

We've seen that it's possible to deduce information about the speech itself by looking at the sound waveform. But it seems obvious that this is not as precise and relevant as seeking information by looking at a text. Speech sounds can be quiet confusing because there's no boundaries between phonemes. The main « issue » that makes speech sounds hard to analyze is that its characteristics can vary a lot, depending on different factors. Theses factors will be detailed below :

- **Physiological Factors:** The size of the vocal tract, which differs between individuals, according to gender and other parameters, influences the pitch frequencies of the voice. Those changes in the pitch cause changes in the waveform of vowels between individuals

- **Behavioural Factors:** The speech can be influenced by behavioural factors such as speaking rate, accent, pronunciation of words depending on the geographical or social backgrounds, etc. Those behavioural factors cause variance between individuals for a same sentence
- **Transducer / Channel:** The transduction is the action of converting the mechanical wave of a sound into an electrical signal (that's what microphones do). This transduction is not done the way with all microphones or recording devices, and that's why it is a factor of variability. For one utterance, the waveform will not be the same if it is recorded with device A or device B.
- **Environmental Factors:** The background noise that is recorded at the same time as the actual speech we want to record influences the signal. That background noise depends on the place and the surrounding environment state at the moment where the sound is recorded.
- **Phonetic Context:** We've seen that we are able to recognize sounds or even phonemes by looking at a spectrum. But it's known that the signal for a sound depends on the sounds that precede and follow this specific sound. The waveform of a sound will be different if it's followed by a sound A or a sound B.

All of these factors lead us to temper our comments when we say that a speech signal is analyzable.

# Chapter 3

# Automatic Speech Recognition (ASR)

## 3.1 What is ASR

Speech is the principal way of communication of the human being. That is why making natural language recognizable and understandable by computers is important and can facilitate the interaction between us and the machine. There can be many applications for an Automatic Speech Recognition (ASR) system like Speech synthesis, Speaker Recognition and Verification, Speech Enhancement, Emotion recognition and many others. Before talking about all these applications, we wwill introduce what is an ASR system. An ASR system have for main objective to transcript speech signal into text in the most accurately and efficiently way and that independently of the recording conditions or environmental noise Also, it would be great if it would be independent of the speaker characteristics (gender, accent, age etc). The difficulties of this task is to handle all the variability of the human voice and to adapt its performance to this variability. We will see the different types of ASR, the difficulties of making a performant ASR then how those type of system work.

## 3.2 Types of ASR

There are several types of ASR for different uses that can describe all the applications of those system and that use particular technologies and system architecture. They can be regrouped in 4 general categories [12] :

- **Number of Speakers :** A Speaker independent system can recognise speech of any speaker usually learned from a very large amount of various data, contrary to Speaker dependent system that perform on a certain type (with certain characteristic) of speaker. Speaker adaptive system work first as a speaker independent system but adapt himself to fit on the actual Speaker to recognize

- **Nature of the Utterance :** In a Isolated Word Recognition system the speaker needs to talk with clear pause between words. It is opposed to Connected Word Recognition system that can recognise words without blank between them. Continuous Speech Recognition handles continuous speech sentences. Spontaneous Speech Recognition manages conversational speech with grammatical error, pause or interjection. Last, Keyword Spotting System that look for specific set of words in the speech input.
- **Vocabulary Size :** A Small Vocabulary system can recognize a small number of words, like for example a keyword spotting system that handle only few words to detect. Medium Vocabulary system can recognise a few hundreds of words. And the Large or Very Large Vocabulary system who are trained with several thousands and tens of thousands of word, used for dictation system for example.
- **Spectral Bandwidth :** Narrow-band speech is a limited channel for example ranging from 300 to 3400 Hz generated by telephone/mobile device where frequency component outside of this range is a lot attenuated. Normal speech that is not affected by this type of techniques limitations is called wide-band speech. A narrow-band speech trained model will have poor performance with wide-band speech data and vice versa, due to the difference in information representation.

## 3.3 Difficulties of ASR

It's been decades since research on ASR started and yet we are still far from achieving human performance. ASR is a decoding process of speech signal and one of the main difficulties come from the variability on the encoding process of this information. It's described as a physiological factor in Section 2.4 and reflect the complexity of the human speech productions. External factors also described in Section 2.4 like environment, behavioral or technical factor add variety to the data and therefore difficulty.

## 3.4 General View

The Figure 3.1 shows the general functionality and process of an ASR system. The input speech signal, passed as a speech waveform $s[n]$, is first converted into sequence of features vectors $X = x_1, \dots, x_T$ by spectral analysis method. Those features are decoded by a pattern classification process that will try to recognized the words, using an acoustic model and a language model that we will talk later, and a word lexicon that represent the word vocabulary. The output is the maximum likelihood sentence translation , given with the confidence score of each word.

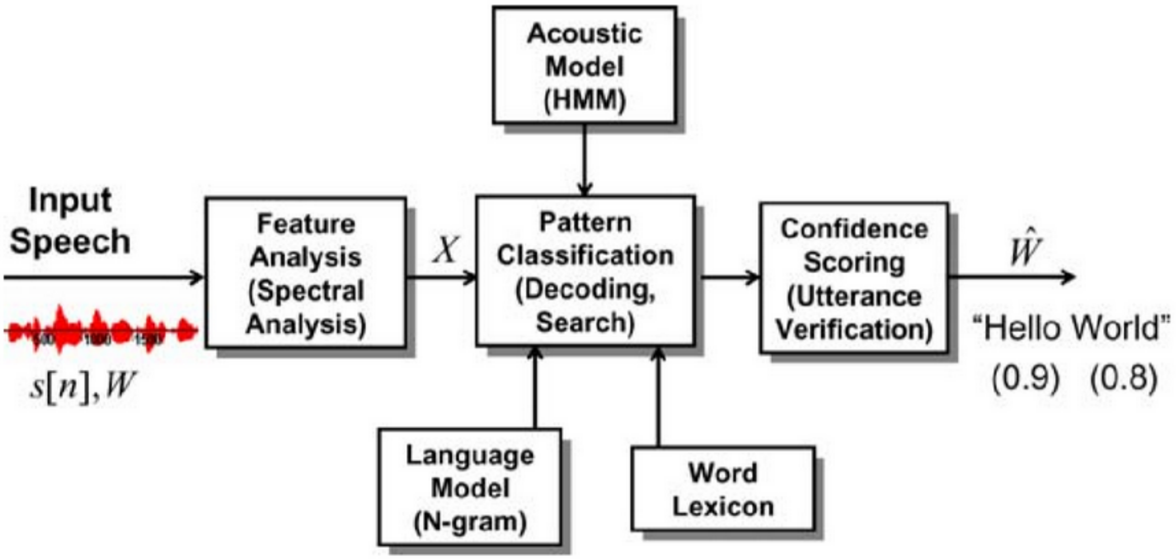
Acoustic Model (HMM)
Input Speech
$s[n], W$
Feature Analysis (Spectral Analysis)
$X$
Pattern Classification (Decoding, Search)
Confidence Scoring (Utterance Verification)
$\hat{W}$
"Hello World"
(0.9) (0.8)
Language Model (N-gram)
Word Lexicon

**Figure 3.1:** Overall Speech Recognition System [11]

We will describe the feature analysis in 3.4.1 and the pattern classification process in 3.4.2 which are the two main process and explain the techniques used in those process in 3.6.

### 3.4.1 Pre-processing / Feature Analysis

Pre-processing is the first step of every speech recognition system, it consist of extracting features from the input waveform Speech Signal that will extract and highlight many information that are good for identifying the linguistic content or some other stuffs like background noise or speaker information, to more easily and more effectively processed the recognition part. There is no typical features to extract, it's an addition of acoustic, articulatory, and auditory features that will depends on the system that is build and the result you want. But throw time of research some extracting method have proven to be efficient. One of the most popular acoustic features is named Mel-Frequency Cepstral Coefficients (MFCC) [8]. They are used to represent the Speech amplitude spectrum in a compact form, to extract the most important information represented on the most optimized way. Figure 3.2 shows the MFCC process steps.

First initial waveform is divide into windows (typically 20ms) using window function like Hamming window that remove edge effects. Then each frame are processed, Discrete Fourier transform is applied to next take the logarithm of the amplitude spectrum that highlight loudness of the signal. Next step is to smooth the 256 spectral actual components into 40 frequency bins equally spaced in frequency. Final step is to apply a transform like Discrete cosine transform to decorrelates those components and obtain 13 cepstral features for each frame.

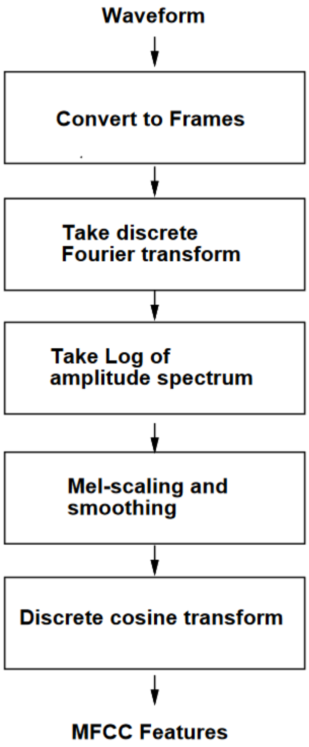
Waveform
Convert to Frames
Take discrete Fourier transform
Take Log of amplitude spectrum
Mel-scaling and smoothing
Discrete cosine transform
MFCC Features

**Figure 3.2:** MFCC features creating process [8]

### 3.4.2 Recognition

The heart of the process is the recognition part that will analysed our features to give the most likely translation of our speech input and its composed of two main training models, the acoustic model and the language model.

**Acoustic Model :** Acoustic model is an important part of the process, it's a statistical representations of all the phonetics unit that makes up word of a language. It's the first step, that recognized the phonemes from a speech signal that will be used by the language model to found the most probable word that is composed by those phonemes.

**Language Model :** Language Model is used after the acoustic model and allow to form sentence that make sense. He can be represented by the grammar rules of a language or simply statistically represent each pair of word, estimated on a training corpus.

Both are estimated or trained on a training labeled data set adapted to the type of ASR where those model will be used.

## 3.5 Performance Evaluation

To improve their performance, any speech recognition system need to have a statistically efficient method to evaluating recognition performance, who is based on an independent test set of labeled utterances. Most of the time, the word error rate and sentence error rate are measured as a recognizer performance. Doing it on an independent set is a very important point, the system must never have encountered the test data in order to not influence and corrupt the results.

## 3.6 Speech Recognition Components

### 3.6.1 Mathematical Formulation of ASR

Automatic Speech recognition problem can be represented as a statistical decision problem [11]. It's a formulation of a Bayes maximum a posteriori (MAP) probability where we want to find a sentence $\hat{W}$ that optimizes the posterior probability $P(W \mid X)$ where $W$ is the input speech to translate, and $X$ is the feature vector of $W$,

$$\hat{W} = arg\,\underset{W}{max}\,P(W \mid X)$$

can be rewrite using Bayes' rule :

$$\hat{W} = arg\,\underset{W}{max}\,\frac{P(W \mid X)P(W)}{P(X)}$$

We can decomposed the precedent equation, and see that the posterior probability is the multiplication of two terms, the prior probability of the word sequence $W$, named $P(W)$, and the likelihood word string $W$ that produced the feature vector $X$, written as $P(X \mid W)$. Knowing that $P(X)$ is independent of the word sequence $W$ which is being optimized, we can ignore it. the acoustic model is represented as $P(X|W)$, commonly denoted as $P_A(X \mid W)$ and the language model is represented as $P(W)$ commonly denoted $P_L(W)$.

$$\hat{W} = \underbrace{arg\,\underset{W}{max}}_{step1}\,\underbrace{P_A(X \mid W)}_{step2}\,\underbrace{P_L(W)}_{step3}$$

We can describe the recognition process as 3 step [11], where Step 1 is the computation of the probability associated with the acoustic model of the speech sounds in the sentence $W$, Step 2 is the computation of the probability associated with the linguistic model of the words in the utterance and Step 3 is the computation of the maximum likelihood sentence resulting of the 2 previous step.

### 3.6.2 Hidden Markov Model (HMM)

The most popular way of generate acoustic models is the used of statistical characterization named Hidden Markov Model (HMM) shown in the Figure 3.3,

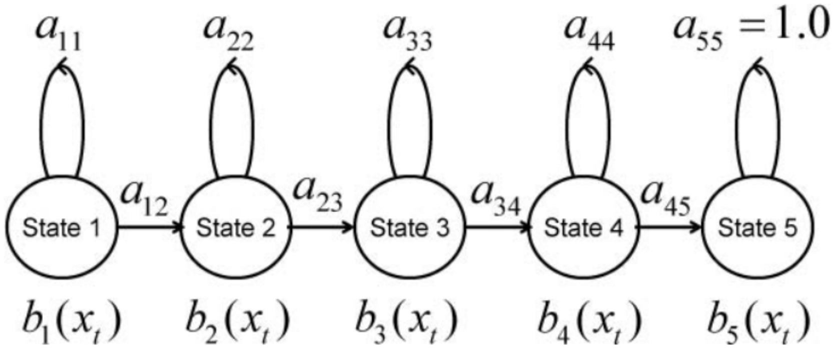

**Figure 3.3:** Q=5 state HMM.[11]

The Figure 3.3 represents a simple $Q = 5$ state HMM used to modeled a word, where each state is characterized by a mixture density Gaussian distribution that represent the statistical behavior of the features vectors within the states of the model. Another characterization of HMM is the state transition $a_{i,j}$ that represent the probabilities of the automata to go from a state to another. Generally the self state transition is high, near to 1 and the state transition is low, near to 0. Another example is shown in the Figure 3.4.

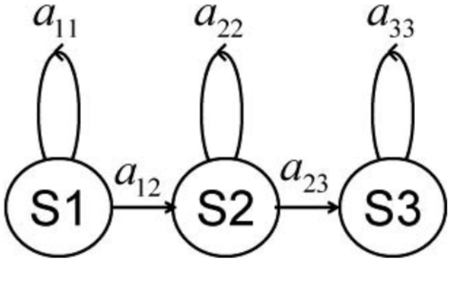

**Figure 3.4:** Q=3 state HMM. [11]

This 3-state HMM can represent a sub-word unit model where the first sate $s_1$ is the statistical representation of the begin of the sound, state $s_2$ the body of the sound and state $s_3$ represent the characteristic at the end of the sound, imagining that this 3-state HMM characterize the sound /HI/, we can concatenate it with another 3 state HMM characterizing the sound /Z/ to model the word "is". Usually, a complete HMM representation of a Q state word model is written as $\lambda(A,B,\pi)$ with $A$, the state transition matrix, $B$, the state observation probability density, and $\pi$ the initial state distribution.

Training of HMM model consist to optimized his model parameters by following a training procedure called Baum-Welch algorithm guided by labeled training set of sentence transcripted as words and/or sub-word units. This algorithm try to align each word (or sub word unit) to the speech input and estimates the appropriate means, covariances and mixture gains for the distributions in each model state, doing it iteratively till a stable alignment is obtained.

### 3.6.3 Gaussian Mixture Model (GMM)

Gaussian mixture models is the most popular way to model HMM state. It's a modeling of the emission probabilities and it can be write as,

$$P_\Lambda(o) = \sum_i c(i)\mathcal{N}(o;\mu(i),\sigma^2(i))$$

where $\sum_i c(i)\mathcal{N}(o;\mu(i),\sigma^2(i))$ is a Gaussian , with $\mu(i)$ the mean, $\sigma^2(i)$ the variance and $c(i)$, the weight of the $i$th Gaussian [7]. More explicitly, GMM is used as a model to represent the distribution of speech features, and are very popular due to their high performance of statistically represent arbitrarily complex distribution from widespread data. Figure 3.5 is a graphical representation of Gaussian mixture in red from the Gaussian components in blue.

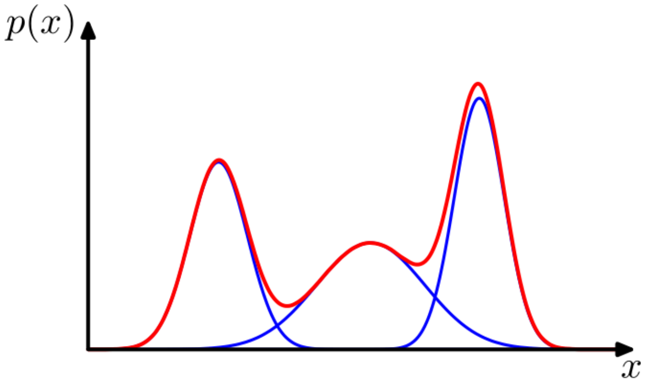

**Figure 3.5:** Gaussian Mixture representation.

### 3.6.4 Deep Neural Networks (DNN)

Deep learning is a branch of machine learning that consist of a set of algorithm that manage learning patterns depending on training data. Deep Neural Network is the most popular model of deep learning. neural network apply to speech recognition research starts during 1980s, and persisted till early 2000s but performance of these method never beat those of GMM-HMM mainly due to the lack of big amount of data that require neural network and deep learning system, and also due to the low computing power of these days (DNN have multiple layers and can go up to a total of 1 million parameters). But it's in the year 2010s that these difficulties have started to been overcome implying an upsurge of popularity for DNN. The further research shown that with a very large amount of data and an adapted architecture of DNN, the error rate drops drastically compared to the best HMM-GMM model, moreover, the error nature of this two method is characteristically different allowing researcher to adapte deep learning to actual highly efficient decoding system. A fundamental principle of deep learning is the used of raw features, system using raw spectrogram shown better result than system using mel-Cepstral features and other features extraction. DNN used in large-scale speech recognition is one of the most successful application of deep learning of this past years. This has made DNN one of the major tools used by all the biggest speech recognition companies today like Apple Siri, Microsoft Cortana, Google now and some others.

# Chapter 4

# Speaker Adaptive Training

## 4.1 What is Speaker Adaptation ?

We've seen that one of the main issue that impacts the accuracy of ASR systems is variability. We also know that there's two types of variability : across-speaker variability and within-speaker variability. The across-speaker variability corresponds to the variability between different speakers. This variability is mainly caused by the factors described in 2.4. Hence, a model trained on, for example, only female subjects will have a bad accuracy if tested on male subjects. The within-speaker variability corresponds to the variability between two utterances from the same speaker. The factors of within-speaker variability can be characteristics such as speech rate, volume, or emotional state of the speaker.

We've also seen the different types of acoustic models that are used to generate ASR systems : Speaker independent acoustic models, which are trained with data from multiple speakers, and speaker dependent acoustic models, which are trained with data from a single target speaker. Those acoustic models suffer from the variability we just discussed. The goal of speaker adaptation is to create acoustic models that are less sensitive to this variability, and therefore improve their accuracy. Applying speaker adaptation methods leads us to create what we call "Speaker Adapted acoustic models" : this type of acoustic model is trained on data from multiple speakers (as a speaker independent model) then is adapted using a smaller set of data from the target speaker (as a speaker dependent model).This smaller set of data is called adaptation data.

In the next part, we're going to take a look at the approaches and methods used to construct these speaker adapted models.

## 4.2 Speaker Adaptation Approaches

### 4.2.1 Spectral Mapping Approach

One approach for speaker adaptation is the spectral mapping approach. In this approach, the goal is to match the new speaker's features vector to the vector of the training data. One way of doing that is using a type of normalization system that maps each new speaker from a Speaker Independent model to a reference speaker so the model acts as if it was a Speaker Dependent model. Basically, the aim is to reduce the differences and the mismatches between the reference speaker data and the new speaker's data.

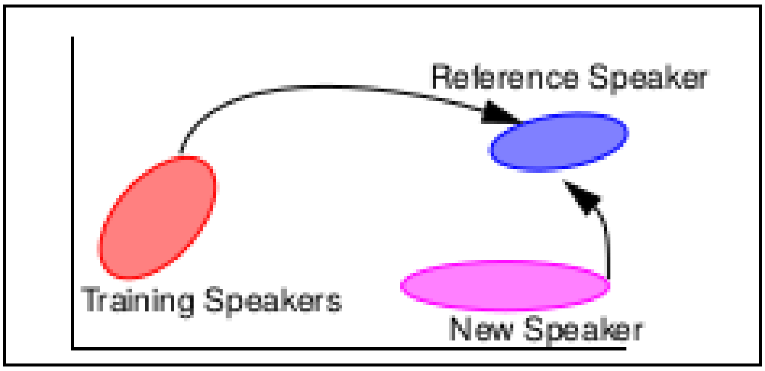

**Figure 4.1:** Graphical representation of spectral mapping [4].

### 4.2.2 Model Mapping Approach

While the spectral mapping approach tries to map all the speakers to one space, model mapping tries to take advantage of the adaptation data in a different way. The model mapping consists in adjusting the model parameters to best represent the new speaker of a Speaker Independent model [4]. For a given speaker, the model tries to adjust to the speaker, by modifying its parameters.

As seen in the figure, model mapping approach uses the adaptation data to transform the parameter of the speaker independent-like model to speaker dependent-like parameters, in order to be closer to the new speaker's features. There are several techniques that are from the model mapping approach, such as the Maximum A Posteriori (MAP) technique or the Maximum Likelihood Linear Regression (MLLR) technique, that we will discuss later. First of all, when talking about model mapping, we have to discuss about the adaptation mode and the training mode, whose properties have an impact on the model.

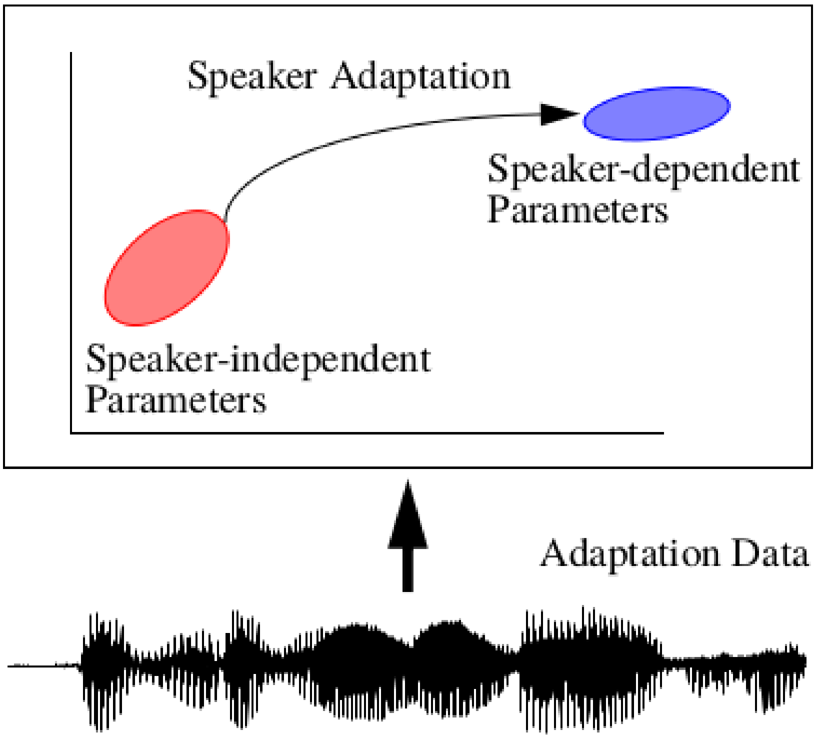

**Figure 4.2:** Graphical representation of model mapping [4].

#### 4.2.2.1 Adaptation Mode : Batch Mode vs On-Line Mode

The adaptation mode is the way the system will use the adaptation data. There are two famous different modes : the batch mode and the on-line mode. With the batch mode, the adaptation data is stored since the beginning of the training, and is used to adapt the model after all the data is collected. With the on-line mode, the adaptation data is incremental. That means that the models are adapted continuously, at each new utterance and their new version are used to produce the next adaptation. We can call this mode “dynamic”.

#### 4.2.2.2 Training Mode : Supervised vs Unsupervised

The training mode is the global way the learning system will work. In supervised learning, the “true” transcriptions of the adaptation data are known. That means these transcriptions are implemented in the system before the training. In unsupervised learning, there are no correct transcriptions available. The system learns how to transcript and feeds itself. Obviously, the supervised learning is more accurate because there’s no transcription error.

## 4.3 Techniques for Speaker Adaptation

### 4.3.1 Maximum A Posteriori (MAP)

To make speaker adaptation we have to estimate parameters. Let's consider a parameter of a probability density function followed by a variable x, describing the model. The distribution of it varies during time and depending on the evolution of the model. During the adaptation, the role of a MAP model [2] is to find a function that will estimate the value of a $\hat{0}$ parameter that will maximize the distribution after the adaptation. The advantage with MAP model is that the accuracy of the Speaker Independent model which is adapted with MAP will converge to the accuracy of a Speaker Dependent model, as the number of adaptation data increases. The disadvantage of the MAP model is that it's not very effective when there's only a little adaptation data. For this type of case, the Maximum Likelihood Linear Regression approach is better.

### 4.3.2 Maximum Likelihood Linear Regression (MLLR)

MLLR is an adaptation method that is more popular than MAP because it can work properly with a small amount of data. As ASR models are based on HMM's, the MLLR adaptation method is applied to to the means and the variances of the Gaussians. The goal is to apply linear transforms to the Speaker Independent model, to adapt it such a way that the likelihood of the adaptation data is maximized [6]. The great advantage of the MLLR method is that a single global transform can carry things out and be used for all models, even with small data.

### 4.3.3 DNN Approach

MAP and MLLR are very popular and efficient method for HMM adaptation and well perform and small amount of data, but such technique does not exit for DNN, they large number of hidden layer involved to many parameters to be optimized, that required to much computing power. Moreover, most of the time, model adaptations result to generate new model for each speaker, which, applying to DNN, increases complexity and requires too much storage. This is why feature adaptation method is the best solution for DNN.

### 4.3.4 fMLLR

Feature-space MLLR (fMLLR) work closely like MLLR (described in Section 4.3.2) but apply to the features input rather than on models, and is known to perform well on small amount of data. This techniques transform acoustic features to speaker adapted feature by multiplying the original features by a transform matrix.

### 4.3.5 I-Vector

I-Vector was recently introduce in ASR and show very high performance on speaker recognition and speaker verification, it quickly become the state of the art on those domain and few research has shown that he can be easily adding to regular feature features to improve DNN performance. The main idea of I-vector is to generate from the speech input one vector of n dimension characterizing the actual speaker identity. The additional in formations bought by I-Vector apply to adaptation techniques is a powerful method increasing efficiency of speech recognition system. Actual research showed that combining I-Vector and fMLLR increases even more the performance.

## 4.4 Summary of planned experiments

Regarding the context of speaker adaptation systems described above, we plan to investigate its effects using several experiments. The related corpus for adaptation is available in the MULTISPEECH team of the INRIA research center. We will be mainly using Librispeech which is large-scale corpus based on English audio books (http://www.openslr.org/12) and Verbmobil which contains multilingual dialogues (https://www.phonetik.uni-muenchen.de/Bas/BasVM2eng.html). Experiments will be performed over both Kaldi speech recognition toolkit (https://kaldi-asr.org) and PyTorch machine learning library (https://pytorch.org).

MFCCs are the most commonly used features, but Perceptual Linear Prediction (PLP) features and other features are also an option. These features will serve as the basis for our acoustic models. In a GMM/HMM framework, a monophone model is an acoustic model that does not include any contextual information about the preceding or following phone. It will be used as a building block for the triphone models, which do make use of contextual information.

The parameters of the acoustic model will be estimated in acoustic training steps; however, the process can be better optimized by cycling through training and alignment phases. These steps will be performed using Kaldi framework. While monophone models simply represent the acoustic parameters of a single phoneme, we know that phonemes will vary considerably depending on their particular context. The triphone models represent a phoneme variant in the context of two other (left and right) phonemes. At this point, we'll also implement triphone units. These will include delta+delta-delta training and speaker adaptive training. The alignment algorithms include speaker independent alignments and FMLLR. Moreover, MAP adapted decoding will be also investigated.

Currently the lattice-free method achieves state-of-the-art results on many speech recognition tasks. This method uses a sentence-level posterior for training the neural network but it is still relies on alignments from a GMM/HMM model. The objective function used in this method is maximum mutual information in the context of hidden Markov models. Following the adaptation context, we will implement these DNN acoustic models to target speakers by supplying i-vectors as input features to the network in parallel with the regular acoustic features for ASR. We will make a detailed comparison of speaker adaptation methodologies at the end of this project.

# Bibliography

[1] Al-Akaidi, M., and Blackledge, J. *Fractal speech processing.* Cambridge university press, 2004.

[2] Gauvain, J.-L., and Lee, C.-H. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE transactions on speech and audio processing 2*, 2 (1994), 291–298.

[3] Gupta, V., Kenny, P., Ouellet, P., and Stafylakis, T. I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2014), IEEE, pp. 6334–6338.

[4] Hamaker, J. E. Mllr: a speaker adaptation technique for lvcsr. *Lecture for a course at ISIP-Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University* (1999).

[5] Karpagavalli, S., and Chandra, E. A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition 9*, 4 (2016), 393–404.

[6] Leggetter, C., and Woodland, P. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. ARPA spoken language technology workshop* (1995), vol. 9, pp. 110–115.

[7] Li, J., Deng, L., Haeb-Umbach, R., and Gong, Y. *Robust automatic speech recognition: a bridge to practical applications.* Academic Press, 2015.

[8] Logan, B., et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR* (2000), vol. 270, pp. 1–11.

[9] Nguyen, T. S., Kilgour, K., Sperber, M., and Waibel, A. Improved speaker adaptation by combining i-vector and fmllr with deep bottleneck networks. In *International Conference on Speech and Computer* (2017), Springer, pp. 417–426.

[10] **Rabiner, L. R., and Juang, B.** A tutorial on hidden markov models. *IEEE ASSP Magazine 3*, 1 (1986), 4–16.

[11] **Rabiner, L. R., Schafer, R. W., et al.** Introduction to digital speech processing. *Foundations and Trends in Signal Processing 1*, 1–2 (2007), 1–194.

[12] **Samudravijaya, K.** Automatic speech recognition. *Tata Institute of Fundamental Research Archives* (2004).

[13] **Saon, G., Soltau, H., Nahamoo, D., and Picheny, M.** Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (2013), IEEE, pp. 55–59.

[14] **Shinoda, K.** Speaker adaptation techniques for automatic speech recognition.

[15] **Tomashenko, N.** *Speaker adaptation of deep neural network acoustic models using Gaussian mixture model framework in automatic speech recognition systems.* PhD thesis, Le Mans, 2017.