ANNÉE 2014

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Stefan ZIEGLER

préparée à l'unité de recherche IRISA – UMR6074

A Study on the Integration of Phonetic Landmarks into Large Vocabulary Continuous Speech Decoding

Thèse soutenue à Rennes le 17/01/2014

devant le jury composé de :

Denis JOUVET
Inria, Loria, Nancy / *Rapporteur*
Martine ADDA-DECKER
CNRS, LPP, Paris / *Rapporteur*
Lori LAMEL
CNRS, LIMSI, Orsay / *Examinatrice*
Laurent MICLET
Univ. Rennes 1, Irisa, Rennes / *Examinateur*
Sebastian STÜKER
KIT, Institute of Anthropomatics, Karlsruhe / *Examinateur*
Guillaume GRAVIER
CNRS, Irisa, Rennes / *Directeur de thèse*

Abstract

This thesis studies the integration of broad phonetic landmarks into standard HMM-based large vocabulary continuous speech recognition (LVCSR). The thesis introduces a general landmark detection framework, that defines landmarks as a sequence of discrete events, indicating the presence of broad phonetic classes in the speech signal. This framework is used to study the two basic issues of landmark-driven speech recognition. The first issue is landmark detection, i.e., the problem of designing landmark detection front-ends which capture relevant phonetic information. Two landmark detection front-ends are presented, which both make use of multi-class classifiers trained on segment-based acoustic observations to classify and detect broad phonetic landmarks in the speech utterance. The second issue is the integration of the obtained landmarks into the search for the best word hypothesis, where two different landmark integration methods are explored: The first method uses binary landmarks as additional pruning criterion and the second method corresponds to weighted combination of phonetic landmarks and the emission probabilities of the acoustic model.

The experimental evaluation shows, that while using phonetic landmarks for pruning is too sensitive towards detection errors and does not improve over standard HMM-based speech recognition, weighted combination of landmarks inside the two proposed detection approaches improves speech recognition on a broadcast news transcription task by 2% relative improvement in both cases. Yet, landmarks do not outperform standard frame-based phonetic classifiers. Since these results indicate that landmark-driven LVCSR might need more heterogeneous landmark models to be effective inside statistical speech recognition, the final part of the thesis presents an extension of the first framework, that attempts to integrate an arbitrary amount of individually designed landmark detection front-ends into the decoding. The proposed framework individually maps each detection function onto a stream of log-likelihood scores, before these scores are discriminatively trained with the acoustic model. Evaluating this framework using detected landmarks of varying accuracy shows that discriminative training leads to many detection functions being essentially discarded during decoding and only those phonetic classes that provide complementary knowledge at frames that are not correctly aligned by the baseline statistical models propagate into improved word hypotheses.

Contents

List of Figures

List of Tables

List of Algorithms

Liste des symboles

ANN	Artificial neural network
AP	Acoustic parameter
AR	Autoregressive
ASR	Automatic speech recognition
BPC	Broad phonetic class
CRF	Conditional random field
DNN	Deep neural network
DP	Dynamic programming
EM	Expectation-maximization
GMM	Gaussian mixture model
HMM	Hidden Markov model
LVCSR	Large vocabulary continuous speech recognition
MFCC	Mel-frequency cepstral coefficient
ML	Maximum likelihood
MLP	Multilayer perceptron
MMI	Maximum mutual information
PDF	Probability density function
SVM	Support vector machine
WER	Word error rate

Introduction

Large vocabulary continuous speech recognition (LVCSR) systems are today at the brink of passing from niche applications and technical gimmicks towards mainstream technology that provides increasing value for customers and companies in a variety of tasks. At first glance, it might seem that this achievement is the result of a rather paradoxical development in the speech recognition community since the beginning of machine-based speech recognition. While early approaches to automatic speech recognition (ASR) relied on joint collaboration of speech scientists and engineers to exploit kowledge about human speech processing for machine-based approaches, ASR made its first leap in performance with the introduction of the statistical ASR framework, which is still the standard approach to ASR nowadays. This statistical framework consists of an acoustic model, usually corresponding to hidden Markov models (HMMs), and a statistical language model, which basically ignore fundamental phonetic and linguistic knowledge about speech and language. Since then, this statistical speech recognition framework has been extended by several important optimization and adaptation techniques that have allowed to efficiently use the ever-increasing amount of training data, up to the point where phonetic knowledge has become nearly irrelevant for the design of modern ASR systems. Despite this still ongoing success story, there are still major limitations of current statistical ASR systems concerning robustness and sensitivity to unseen speech data. Since humans largely outperform machine-based ASR on complex recognition tasks, many studies claim that the ignorance towards phonetic knowledge of modern ASR is a bottleneck that prevents automatic speech recognition to finally reach human-like recognition accuracy.

The difficult relation between phonetic knowledge and ASR

It is one thing to point out the obvious shortcomings of current ASR systems compared to human speech recognition, but a whole other thing to overcome the numerous challenges that come along with changing the acoustic modeling from a phonetically ignorant approach towards a modeling paradigm that accommodates for phonetic knowledge without loss in recognition accuracy. Indeed, taking a closer look at the body of scientific work on human speech production and perception from the last decades might lead to more questions than answers for speech engineers, since the linguistic community had its very own shift in paradigms. Theories about human speech perception changed from what was supposed to be a linear mapping of invariant cues onto discrete phonological elements, to a complex non-linear process that is supposed to operate on several layers with various kinds of information transfer in-between. While there is a growing body of experimental and theoretical results that explain specific speech phenomena and show relations between acoustic content and perceived speech units under certain conditions, this knowledge mostly remains partial knowledge and the complexity of the problem prevents parallel existing theories about speech perception to converge into one general model. Thus, alternative acoustic modeling techniques that have been proposed over time have to fill the black boxes of phonetic theories and struggle at times to find the right balance between more complex models, which are difficult to train and make the search for the best word hypothesis intractable or phonetically motivated heuristics, which often lack generalization and the possibilities for mathematical optimization.

Towards landmark-driven ASR

The goal of this thesis is to modify the state-of-the-art statistical ASR decoding framework, to accommodate for phonetic knowledge by combining standard HMM-based decoding with landmark-based approaches to ASR. Landmark-based approaches to ASR rely on detecting selected time instances as landmarks in the speech utterance which indicate the most salient points of articulatory gestures. Evaluating acoustic information in vicinity of these potentially perceptually relevant points gives evidence about higher level speech units. In this thesis, landmark detection functions indicating the presence of phonetic classes are used to guide the decoding of a standard statistical ASR system according to the detected phonetic information if a landmark is present in the speech signal and to perform standard Viterbi decoding if not.

While using landmarks as the sole acoustic modeling paradigm is far from competing with the statistical power of HMM-based ASR, they possess several attributes which make them a promising complementary model to standard HMMs. First, there is a considerable amount of freedom in adapting phonetic detection front-ends according to phonetic knowledge, which in return might provide complementary acoustic information to the standard emission probabilities of the acoustic model. Second, landmarks can afford to concentrate on those phonetic classes which are associated with well studied acoustic cues and model only those parts of speech which show relatively few ambiguity about the acoustic content.

This thesis addresses the two basic issues that have to be solved on the road towards landmark-driven ASR. First, there is the issue of converting the abstract concept of phonetic landmarks into detection front-ends, including choosing which phonetic classes are reliable to detect and examine how to determine perceptual relevant landmarks in the speech utterance. The second issue is the integration process, i.e., the question of how to modify the standard Viterbi decoding according to the detected phonetic information.

Contributions

The thesis introduces a general landmark detection framework that consists of a bank of landmark detectors, each detector indicating the presence of a natural speech class, for example broad phonetic classes vowels, sonorants, plosives and fricatives. These detection functions are converted into binary landmarks that are used to bias the search space of a standard HMM-based speech recognizer.

This framework is used to extend the experiments conducted in a previous study ([GM07]) concerning a landmark-based pruning method that prunes the search space during decoding according to detected broad phonetic landmarks. While the method in [GM07] achieved considerable reduction of the word error rate using oracle landmarks, the follow-up experiments in this thesis use landmarks that are based on statistical classifiers and thus are subject to realistic acoustic confusions. The results indicate that detection errors have indeed a detrimental effect on the final word hypothesis and reducing phonetic confusions by detecting very broad phonetic classes does not provide complementary information for the baseline system. The results from these experiments lead to changing the integration method towards a weighted combination of binary landmarks and acoustic scores, to bias the search space according to phonetic information without completely pruning paths. This method is used in the first and second contribution of this thesis, corresponding to two landmark detection approaches that propose different landmark detection strategies.

The first approach relies on a segmentation and classification approach, i.e., speech is segmented according to estimated articulatory movements and the obtained segments are used to extract acoustic observations for each segment. Training broad phonetic classifiers on this segment-based observations leads to an improved classification accuracy, compared to standard frame-based phonetic predictions on a phonetic detection task.

The second approach focuses on extracting fixed-dimensional acoustic observation vectors from time-variable speech units. The acoustic content of a time-variable speech segment is reduced to spectral homogeneous subsegments and mapped onto a fixed-dimensional observation vector that is used to train segment-based classifiers. These classifiers are used to produce a frame-based detection profile by searching for each frame the corresponding segment that has the highest classification score, which is used to detect landmarks as local maxima of the obtained detection profile.

While both approaches achieve a small, but significant, improvement on a broadcast news transcription task, experiments show that the improvement can equally be obtained by training broad phonetic classifiers on regular continuous frame-based observations. The major reason for this result is supposed to be found in the use of a shared front-end for all phonetic classes that relies on homogeneous acoustic observations for all phonetic classes. In the end, these landmarks can not provide complementary acoustic information to the emission probabilities of the HMM-based acoustic model.

The third contribution is motivated by these results and provides a new integration framework that attempts to integrate individually designed landmark detection front-ends into standard ASR. In the proposed framework, the landmarks obtained by each detector are individually mapped onto a log-likelihood, before the obtained likelihoods are jointly trained with the emission probabilities using discriminating training. The experimental evaluation shows that speech recognition improves, if landmarks provide complementary knowledge at frames that are not correctly aligned by the baseline statistical models.

Outline

The first chapter introduces into the basics of phonetics and phonology, with emphasis on phonetic studies on human speech perception and the second chapter presents the state-of-the-art statistical ASR framework. The third chapter explains the reasoning behind choosing to integrate phonetic landmarks into HMM-based ASR by reviewing alternative acoustic modeling approaches and introduces the concept of phonetic landmarks. The major contributions of this thesis begin at chapter four, with the introduction of the general landmark detection framework, with emphasis on two different general architectures: one consisting of an individual front-end for each phonetic class, which allows a heterogeneous modeling approach, and one shared front-end for all phonetic classes, which leads to homogeneous phonetic modeling. This chapter also introduces the two methods used for integration of landmarks into the decoding in this thesis, which are classifier combination and landmark-based pruning. Chapter five extends existing studies on the use of phonetic landmarks for pruning the search space during decoding, which motivates the weighted combination of landmarks and emission probabilities, which is first presented in chapter five. Chapters six and seven present the two different landmark detection frameworks that have been developed during this thesis, one based on a segmentation and classification framework, the other discussing the use of segment-based classifiers to obtain a frame-based detection profile. Chapter eight concludes the contributions of this thesis and presents a new landmark integration framework that allows to integrate heterogeneous and asynchronous landmarks resulting from individually designed landmark detection front-ends into the decoding of statistical ASR. The thesis concludes with a short summary and an outlook on future work.

1. Phonetic and phonological basics of speech

This chapter briefly introduces two major subfields of linguistics: phonetics and phonology. The field of phonetics includes all studies related to human speech production, the acoustic properties of speech sounds and their perceptual effects. Phonology is concerned with organizing the different sounds of a particular language into abstract representations and studies the relation between acoustic contrasts and meaningful speech elements.

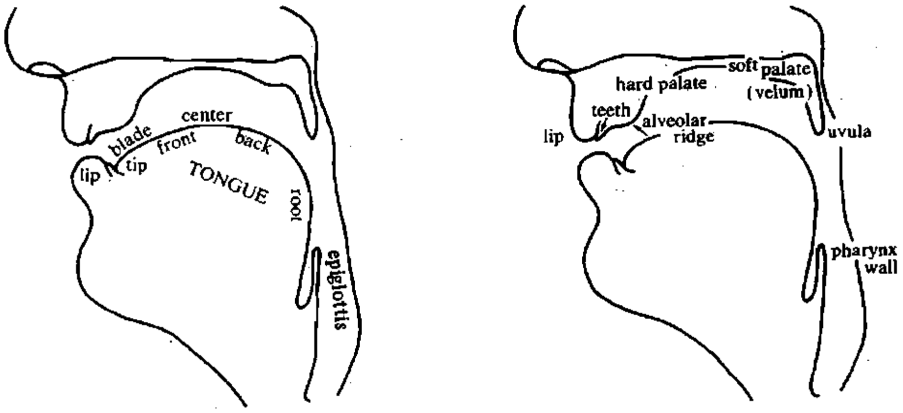
The purpose of this chapter is to give an overview on how speech scientists approach the fundamental questions on human speech processing: How do humans encode information into the acoustic signal? How is the perceived signal converted into linguistic units? What are the elementary building blocks of speech? What is the relation between acoustic content and different types of speech representation? While it will be seen that many of these questions still remain open questions nowadays, the results obtained from decades of theoretical and practical research in the field of phonetics and phonology, that are briefly summarized in this chapter, can reveal obvious limitations of modern machine-based approaches to speech recognition. Furthermore, phonetic knowledge can point towards concepts of human speech production and perception that are sufficiently understood to serve as basis for developing new models for speech recognition.

1.1. Speech production

Human speech production is often approximated by the source-filter model of human speech. In this basic model, a source signal is transformed by an acoustic filter into the acoustic waveform that is perceived as human speech by a listener (see [Fan71]) .

The source The basic source of human speech is the airstream that is provided by the lungs during respiration. If this air stream causes the larynx to vibrate, the source corresponds to a periodic excitation signal, referred to as *voiced* speech. If the air stream passes through the open glottis without periodic excitation, the produced speech sounds are considered to be *unvoiced.*

The filter In a simple model, the physical properties of the vocal tract, along with the articulators that are involved in speech production, can be considered as a time varying acoustic filter that transforms the excitation signal into intelligible speech. The vocal tract includes the oral as well as nasal cavities and the principle organs of articulation involved in speech production are displayed in Figure 1.1.



(a) lower surface

(b) upper surface

Figure 1.1.: Lower and upper surface of the vocal tract and the principle articulators [Lad82].

1.2. Classification of speech sounds

The type of sounds that can be produced by a human speaker are constrained by the physical limitations of the organs involved in speech production. The basic speech sounds that can be produced by humans are referred to as *phones.* Each phone is produced by a different articulatory gesture, resulting in different acoustic and perceptual properties for each phone. Thus, a speaker can convert an intended utterance into a sequence of acoustically distinctive elements during speech production and a listener consequently decodes the continuous acoustic signal to obtain the original message. Phones can be categorized into a variety of overlapping categories, usually according to common articulatory or acoustic properties. In additional to the source, which divides speech sounds into *voiced* and *unvoiced*, the basic dimensions that are used to categorize speech sounds are *manner of articulation* and *place of articulation*, which are further described in the following paragraphs. Table 1.1 distributes the phone inventory[1] of the baseline speech recognizer used in this thesis into the main phonetic categories of the International Phonetic Alphabet (IPA) [Int99].

Manner of articulation The most basic way to produce several distinctive speech sounds is to simply vary the degree of closure that constricts the airflow passing through the articulatory system. A common way to classify the various types of constrictions is to distinguish between five fundamental articulatory gestures: vowels, approximants, fricatives, plosives and nasals, commonly referred to as broad phonetic classes (BPCs). *Vowels* are sounds produced without constricting the airflow, while the remaining sounds

[1] The phonetic alphabet used in this thesis corresponds to the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) [GMM00], which is a mapping of the phonetic labels of the common IPA notation onto 7-bit ASCII characters.

place → manner ↓	labial	dental	palato-alveolar	palatal	velar	uvular
nasal	m	n		J	N	
plosive	$p\ b$	$t\ d$			$k\ g$	
fricative	$f\ v$	$s\ z$	$S\ Z$			
approximant		l		$j\ H$	w	R

(a) consonants (with manner of articulation on the y-axis and place of articulation on the x-axis)

backness →	front		central	back
height ↓	unrounded	rounded		
close	i	y		$u\ U \sim$
	e	2		
mid	E	9		$o\ o \sim O$
open			$a\ a \sim$	

(b) vowels (with vowel height on the y-axis and vowel backness on the x-axis)

Table 1.1.: Simplified X-SAMPA consonant and vowel charts of French phonemes according to [Can05]. Row and columns group the phonemes into different classes that share a common acoustic and articulatory property. The phonemes displayed are limited to the phone inventory of the baseline speech recognition system used in this thesis (see chapter A.3). The symbol «∼» indicates nasalized vowels.

are all produced by narrowing the airflow through the vocal tract, referred to as *consonants*. *Approximants*, also referred to as *semi-vowels*, are consonants that are similar to vowels in the sense that they do not include any turbulent airflow. Nevertheless, the airflow is significantly narrowed by the vocal tract during pronunciation. If the constriction is narrow enough to produce turbulent sounds, one refers to these sounds as *fricatives*, while producing a speech sound by creating a full closure of the vocal tract corresponds to a *plosive*. *Nasals* are also produced by a full closure of the vocal tract, but the air escapes freely through the nasal cavity.

Place of articulation While manner of articulation describes the type of constriction, place of articulation specifies where in the vocal tract the constriction occurs and which articulators are involved in creating this constriction. Consequently, manner of articulation only concerns consonant speech sounds. Figure 1.1 displays the various articulators defining the places of articulation inside the vocal tract and Table 1.1a lists the main phonetic categories according to the place of articulation.

Vowels With place of articulation categorizing consonants, the different types of vowels are usually described by four general dimensions that shape the acoustic filter during articulation: vertical position of the tongue (height), position of the tongue relative to the back of the oral cavity (backness), the shape of the lips (roundedness) and whether the air stream is partly released through the nasal cavity (nasalization).

1.3. Speech perception

Theories about human speech perception have undergone a significant amount of changes during the last century, since initial theories have been more and more rejected by experimental and theoretical studies in phonetics, linguistics and cognitive psychology.

1.3.1. Characteristics of humans speech - what makes human speech special?

Given the human ability to perceive a continuous acoustic signal and converting it into a sequence of discrete phonological and semantic units, for example a sequence of phones or words, early studies on human speech perception (e.g., [Lic52]) assumed that human speech perception is based on extracting invariant acoustic properties from a linear segmentation of the speech utterance. Backed up by perceptual and neurophysiological experiments, modern theories about speech perception generally dismiss this traditional view on speech, since this view is not able to account for many basic characteristics of human speech perception ([GP88]). It is important to compare the traditional and modern view of human speech perception, since modern state-of-the art ASR largely follows the outdated traditional view on speech, which helps to explain some of the limitations that state-of-the-art ASR has been attempting to overcome during the last decades. The following paragraphs briefly compares traditional and modern views on human speech perception according to [PL07].

Linearity of speech

Traditional view: Since speech is perceived as successive discrete linguistic units, like phones or words, the acoustic information necessary to identify each element is to be be found in ordered non-overlapping successive portions of the signal.
Modern view: Several perceptual experiments have successfully demonstrated that while humans perceive speech as a linear sequence of discrete linguistic units, each time instance in the acoustic signal can carry information about present, preceding and following phones. For example, it has been shown that the length of a vowel is important for perceiving a following consonant as voiced or unvoiced [Lis86].

Speech segmentation

Traditional view: Each higher order linguistic unit, like words or phones, can be assigned to a temporal variable segment in the waveform and an utterance can effectively be sliced into a sequence of speech sounds, that can be recombined to form new utterances.
Modern view: Speech theories nowadays agree that there are no clear acoustic, phonetic or perceptual boundaries between successive speech units and the particular acoustic realization of a speech unit heavily depends on the context.

Invariance of speech sounds
Traditional view: Each of the segments associated with a phone possesses invariant acoustic properties, that makes them identifiable regardless of the context, speaker and transmission channel.
Modern view: While experiments have identified several basic acoustic properties that can be associated with the perception of certain speech sounds under controlled conditions, it is unclear whether a speech unit can be linked to a fixed set of cues that have to be preserved in the acoustic signal to enable the perception of this speech unit. Generally, the mapping from acoustic attributes to discrete linguistic units has shown to be very complex and is not understood yet.

The main reason for rejecting the traditional view on speech is the fact that the associated theories are not able to explain speech perception in the light of the main sources of speech variability, which are summarized in the next paragraph.

Sources of variability in speech Despite the lack of invariant acoustic properties of the speech signal, humans can account for several sources of speech variability that do not diminish the human ability to correctly perceive spoken utterances:

- *Inter-speaker variability* summarizes the differences in the acoustic signal due to different speaker identities. Obvious reasons for inter-speaker variability are individual physiological constraints of the vocal tract, changes in the speaking style and the social or ethnic background of a speaker.

- Even one phrase uttered by the same speaker can result in very different acoustic waveforms. Speech always transmits information that goes far beyond the raw word sequence. Emotion, health and the environment influence the speaking style of a speaker, which adds *intra-speaker variability* to the speech signal.

- Changes in the *acoustic channel*, which corresponds to a filter transforming the input signal into the perceived waveform, do, to a certain degree, not affect human speech perception, while significantly transforming the input signal. Acoustic channels can range from different acoustic environments to narrowband speech transmissions.

- Due to the inertia of the articulators, articulatory gestures change with respect to the preceding or following phonological units, especially in connection with rapid speaking styles. This leads to *coarticulation*, where an articulatory movement is not fully carried out, but the articulatory target is nevertheless perceived as present due to its context.

Despite this variability in the signal, humans are able to distinguish different speech sounds by detecting acoustic cues in the speech signal. While this detection process is effortless for humans, identifying the relevant acoustic cues in the spectrogram is a difficult task for speech scientists.

context	
during closure	duration of closure, duration of glottal signal, intensity of glottal signal
pre-closure	duration of vowel, duration of first formant transition, first formant offset frequency, first formant transition offset time, timing of voice offset, fundamental frequency, decay time of signal
post-closure	release burst intensity, timing of voice onset, onset of first-formant transition, first formant onset frequency, first-formant transition duration, fundamental frequency contour

Table 1.2.: Acoustic cues for voiced/unvoiced distinction of plosives according to [Lis86] for the minimal pair «rabbit» and «rapid». Cues are grouped according to whether they can be found during the closure of the plosive, before the closure or after the closure. It is important to note that the ensemble of cues presented in this table are only valid for plosives in this specific context, i.e., placed between the two vowels of the minimal pair. In a different context, plosives might be detected by a different set of cues.

Acoustic cues The search for the cues in the acoustic signal that make linguistic units distinguishable and trigger the perception of a certain speech unit is an ongoing field of research. In speech science the term «acoustic cue» usually follows closely the definition in [Rep82]:

> A cue [...] is a portion of the signal that can be isolated visually, that can be manipulated independently in a speech synthesizer constructed for that purpose, and that can be shown to have some perceptual effect.

Acoustic cues can be described by the three dimensions of information that are present in the spectrogram of a speech sound: time, frequency and intensity. The mapping from acoustic cues to the perceived linguistic unit is effectively a many-to-many mapping, i.e., one acoustic cue can trigger the perception of several different linguistic units, while one linguistic unit can be caused by different sets of acoustic cues depending on the context. Therefore, the search for the relevant acoustic cues and their interactions among each other has to take place in tedious and costly cognitive experiments. To give an impression on the spectral and temporal variety, as well as complexity of acoustic cues, Table 1.2 summarizes the acoustic cues that have been found to be relevant for the distinction of the two minimal pairs «rabid» and «rapid», i.e., the distinction between voiced and unvoiced stop consonant.

1.3.2. Models of human speech perception

While many studies have provided valuable insights into the process of speech perception, science has not converged to one single theory about human speech perception yet.

The following paragraphs will summarize some major themes of widely acknowledged theories of human speech perception, without organizing them into a global overview.

Hierarchical processing Most theories about human speech perception (for example the TRACE model of speech perception [ME86]) account for several hierarchical processing stages that process the raw acoustic signal to intermediate representations before the brain perceives a sequence of phones and finally words and meaning. Theories differ in the amount of information exchange that takes place between the individual stages and the degree to which lexical and semantic knowledge influence the early stages of human speech processing.

Parallel processing and multiple information streams Since speech perception requires the processing of huge amounts of information, the human brain is supposed to process and extract information from multiple streams in parallel [CLA05]. Early processing stages are supposed to be composed of a bank of several auditory detection mechanisms that are specialized in detecting perceptual relevant cues, before this information is connected to form higher level speech units.

Acoustic pattern matching and speech variability All theories of speech perception attempt to explain the process behind the human ability to recognize speech sounds despite the huge variability they are exposed to. The motor theory of speech [OM78] emphasizes the importance of articulatory gestures for human speech perception, by directly linking the human ability of speech perception to the human knowledge of speech production. In this theory, articulatory gestures are stored in the human brain as invariant gestures that can be recovered from analyzing the acoustic signal. The fuzzy logical model [LCSSK67] models speech perception as hierarchical process that matches stored prototypes against acoustic content by a combination of logical rules.

Despite the lack of an unifying theory on the mapping of acoustic content to linguistic units, phonological studies and of course automatic approaches to speech recognition, work with different forms of discrete speech representations. The next section introduces into the different levels of speech transcription and how speech labels are related to the acoustic signal.

1.4. Transcribing speech and speech representations

Speech transcription is the task of attributing discrete labels from a given set of speech labels to time events or temporal segments in the speech signal. The labels can define speech regarding physical, acoustic or higher level linguistic properties. The task of speech labeling is often very loosely defined and usually subjective to a particular task for which it is produced.

The goal of this section is to get an overview on the type of speech labels that exist and how they can be produced. This is important in order to realistically estimate the limits of machine-based approaches to speech recognition that use these transcriptions to map the acoustic signal to discrete labels.

The different levels of speech transcription presented in the following, including physical, acoustic-phonetic and citation-phonemic level, correspond to a compressed version of the classification presented in [BF92] which overlaps with previous (e.g., [APR89, HAB$^+$90]) and following (e.g., [HKT95]) work in that field. Visual examples for each of the following levels of transcription taken from [BF92] can be found in Section A.1 in the Appendix.

Physical level At the physical level, each label is associated with a physical property of the speech signal that is supposed to be perceptually relevant. There are few constraints on the type of label, comprising discrete time instances and (potentially overlapping) time segments. Additionally to the time, physical labels can make use of frequency and intensity of the spectral representation to provide a very accurate description of the physical event. Physical labels are the result of measurement operations, including subjective acoustic measurements, visual measurements of the spectrogram and various forms of actual physical measurements like (electro-) palatography. Usually, labels are placed according to strict rules and placing labels involves strict parameters, like a set of thresholds, that convert continuous measurements into discrete categories.

Acoustic-phonetic level Acoustic-phonetic labels are associated with speech events that have a phonetic relevance. While these events are mostly associated with discrete time events, they can also stretch over a temporal segment in the speech signal. Labeling has to be carried out by an expert, that has profound knowledge about relevant phonetic events in the speech signal. The labels can become rather subjective when they involve the setting of boundaries of segment-based speech events and decisions about weakly articulated phonetic events. Phonetic events are often labeled with regards to their relevance for higher-level transcriptions, like the following narrow phonetic level.

Narrow phonetic and citation-phonemic level Narrow phonetic transcriptions attach the labels of the IPA to time-variable segments. The segment attached to a label should contain the major acoustic cues of the perceived speech sound. Since speech is per se not suitable for being sliced into successive segments, narrow phonetic labeling is always subjective and needs elaborated labeling guidelines, to determine which parts of the signal correspond to which label.

Phonemic labels are similar to narrow phonetic labels, with the important difference that the speech labels only contain the phone inventory of a given language, instead of the quasi-complete set of IPA symbols. It is referred to as citation-phonemic, since the pronunciation of a word is determined by the citation form of a word that is present in a lexicon. Since these pronunciations correspond to ideal pronunciations, determined by comparing several carefully pronounced realizations of a word, the actual acoustic realization might differ considerably from this canonical representation.

It should be noted that citation-phonemic labeling is an analytic concept that is

b	$O\sim$	Z	u	R
$\# - b + O \sim$	$b - O \sim + Z$	$O \sim - Z + u$	$Z - u + R$	$u - R + \#$

Figure 1.2.: Citation-phonemic transcription of the French word «bonjour» using context-independent phones (upper row) and context-dependent phones (lower row). # is signaling the context-free beginning and ending of the word.

useful for analyzing speech, especially in the context of technical speech applications. The labels are meant to be a mediator between the physical signal and the ideal lexical representations of words but do not describe the mapping of acoustic content onto perceived linguistic units, which is non-linear and essentially not understood yet.

ASR and speech transcription Speech transcription in ASR follows the citation-phonemic labeling paradigm, i.e., the speech signal is divided into a sequence of temporal segments, with each segment being attached with a basic speech unit, usually a *phoneme*. A phoneme is a minimal unit, i.e., changing from one phoneme into another changes the meaning of a word. In contrast to phones, phonemes are abstract units of a language and do not refer to actual physical segments, like phones. Thus, several acoustically distinct phones can be summarized under the same phoneme if the perceived meaning is identical, referred to as *allophones*. Since the pronunciation of speech sounds varies depending on the context, state-of-the-art ASR systems use context dependent phones (*triphones*) as speech units, i.e., each phone is considered in the context of its preceding and following speech units (see Figure 1.2).

Distinctive features Distinctive feature systems [CH68], also referred to as phonetic or articulatory feature systems, group phoneme inventories into natural speech classes according to their articulatory and acoustic similarities. The atomic speech units of the distinctive feature alphabet correspond to binary features, with one feature describing a particular property that will either be true or absent for a certain phoneme. Table 1.1 can be converted into a simple feature system by associating each row and each column with a phonetic class, corresponding to a distinctive feature, that will be «true», respectively «false» for the phonemes of this row (or column), and the inverse for all remaining phonemes.

1.5. Summary and conclusions

This chapter first introduced human speech production with explaining the source-filter model of speech and the basic speech sounds. The second part concentrated on human speech perception, emphasizing the possibilities and limits of speech science in formulating models of speech perception. The chapter concluded with an introduction into the different levels of speech transcription and introduced phonemes and distinctive features as two examples for speech label alphabets. The remainder of this thesis will focus on detection and classification of phonetic classes that share common acoustic properties, i.e., can be described by a fixed set of binary distinctive features. While the main focus will lie on broad phonetic classes, i.e., manner of articulation, some methods will be easily extendable to account for other phonetic classes derived from distinctive features.

Three points are important to keep in mind for the remaining chapters. First, there is no single theory about speech perception, so that incorporating elements of a specific theory into a machine-based approach for speech recognition always follows a subjective view on speech, which does not account for all speech phenomena. Second, while there is partial knowledge about the mapping of acoustic patterns and perceived phonological unit in controlled conditions, the acoustic cues involved in this mechanism are too variable to allow reliable prediction of perceived speech units from the acoustic signal in uncontrolled speech environments. Third, citation-phonemic speech transcriptions, as they are used in machine-based ASR, cannot provide detailed information about the relation of perceived mental speech units and physical signal, since they can only show the abstract relation between acoustic signal and «ideal» pronunciations obtained from a dictionary.

2. Statistical speech recognition

This chapter introduces the statistical formulation of the speech recognition problem and summarizes the architecture of modern statistical ASR, with emphasis on the components that are particularly relevant for this thesis: HMM-based acoustic modeling and the basics of search in large vocabulary ASR.

Pioneers of modern statistical speech recognition include the works of Baker [Bak79], Jelinek [Jel97] and Rabiner [Rab89, RJ93] which introduced the general speech recognition framework that is still valid today, i.e., HMMs as statistical tool for acoustic modeling and n-grams or context-free grammars for the probabilistic modeling of word sequences.

2.1. Problem formulation

The problem of speech recognition consists in searching a sequence of words $\hat{W} = \hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n$ given an acoustic observation X. The hypothesis $\hat{W}$ is ideally equal or at least very similar to the originally uttered word sequence, that has been encoded in X. The fundamental equation of statistical speech recognition uses Bayes' decision rule to search the word sequence $\hat{W}$ that maximizes the posterior probability $p\,(W \mid X)$ among all possible word sequences,

$$\begin{aligned} \hat{W} &= \arg\max_{W} p\,(W \mid X) \\ &= \arg\max_{W} p\,(X \mid W)\, p\,(W)\,. \end{aligned} \tag{2.1}$$

Equation 2.1 incorporates the two basic knowledge sources a statistical speech recognizer has to provide:

- The *acoustic model* captures knowledge about the mapping from words to acoustic observations for calculating $p\,(X \mid W)$.

- The *language model* incorporates statistical knowledge about word sequences to provide the a priori probability $p\,(W)$ of a word sequence W.

This general statistical approach to ASR can be applied to a number of different scenarios ranging from the recognition of isolated words to large vocabulary continuous speech recognition (LVCSR). The following section will give a general overview of the architecture of state-of-the-art LVCSR architectures.

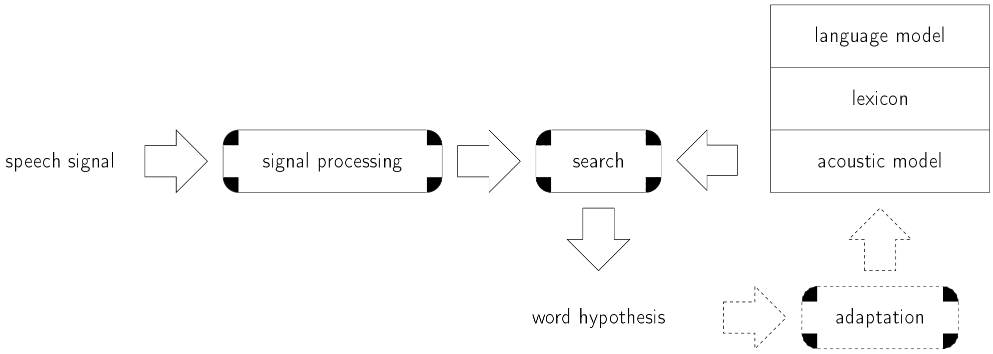


Figure 2.1.: System components of modern LVCSR recognizers.

2.2. Architecture of modern LVCSR

Figure 2.1 displays the main components of a state-of-the-art large vocabulary speech recognizer, including the signal processing stage that converts the acoustic speech signal into a parametrized speech representation and the decoding stage that uses acoustic, language and lexical knowledge to search the most likely word hypothesis. There are various forms of acoustic and linguistic adaption techniques that can additionally be used to refine the hypothesis in multiple passes.

2.2.1. Speech parametrization and pre-processing

The first step in ASR is to convert the acoustic signal into a sequence of t k-dimensional observation vectors,

$$X = x_1, x_2, \ldots, x_t. \tag{2.2}$$

Each vector x_t corresponds to the parametrized speech waveform inside a small observation window of about 10-30 ms length, where the speech signal is assumed to be stationary. Conventional spectral representations like short-time Fourier transform or short-time discrete cosine transform are not suitable for speech recognition since the coefficients of such spectral vectors are highly correlated, to a large degree speaker and channel dependent and do not take into account the human auditory system. Therefore, the coefficients resulting from short-time spectral analysis are further processed into compact representations like the common mel-frequency cepstral coefficients (MFCCs) [DM80] or perceptual linear prediction (PLP) coefficients [Her90].

Additional pre-processing steps involve channel and speaker equalization, for example by cepstral mean subtraction (CMS) or cepstral variance normalization (CVN). Single speech frames are normally enhanced by context information using first and second order derivatives or concatenated super-vectors that span typically over 100ms and are mapped to a lower dimensional space by linear discriminant analysis (LDA) or heteroscedastic LDA (HLDA). Features can be further transformed, for example to reduce the inter-speaker variability by vocal tract length normalization (VTLN), or can be discriminatively trained by neural networks to add bottleneck features to conventional feature vectors [GKKC07].

2.2.2. Acoustic modeling using hidden Markov models

Hidden Markov models (HMMs) have been the quasi-standard for acoustic modeling for several decades. While there have been several promising alternatives throughout the years, like Dynamic Bayesian networks [ZR98] or conditional random fields [ZN09], no technique was able to compete with conventional HMMs and their various form of adaptation techniques. This was the case until recent years, when the use of computationally much more expensive deep neural networks (DNNs) has started to outperform the classic continuous density HMMs on a variety of tasks (see for example [DYDA12, SLY11]).

Yet, it is still valid to refer to HMMs as the state-of-the-art technique for acoustic modeling, since DNNs rely on the existing HMM models to train their networks and do not provide a complete training and prediction framework. Indeed, DNNs do just replace one component of classical HMMs, by using neural networks instead of GMMs to predict the state emission probabilities. Thus, they do not provide a new paradigm for acoustic modeling, but are essentially refining the existing HMM framework and do not overcome basic structural limitations of HMMs.

HMMs as finite state automata An HMM is a finite state automaton, i.e., a set of states $S = \{s_1, \ldots, s_i\}$ connected by arcs that correspond to possible transitions between the states. Each state models a region of the acoustic space, spanned by the acoustic observations X, using a probability density function (PDF) and the transition between states models the temporal evolution of the state sequence. Thus, the emission probability $p\left(x_t \mid s_i\right)$ of a state s_i corresponds to the probability that the acoustic observation at t has been emitted by state s_i.

Topology of HMMs The set of states S, together with their initial probabilities and the state-transition matrix A, determines the topology of a HMM. Given a set of emitting states $S = \{s_1, \ldots, s_i\}$, the transition matrix is an $i \times i$ matrix A, with element a_{ij} containing the transition probability from state s_i to state s_j. For state s_i, all possible transitions sum up to 1, with $\sum_j a_{ij} = 1$. The i-th initial probability $p\left(s_i\right)$, with $\sum_i p\left(s_i\right) = 1$ corresponds to the probability that state s_i will be the first state of the observed sequence. A common topology for a HMM in ASR is a three-state left-to-right HMM, which consists of three emitting states with only one entry state with $p\left(s_1\right) = 1$ and one exit state.

Continuous density HMM The emission probabilities $p\left(x_t \mid s_i\right)$ are usually modeled by continuous PDFs, typically Gaussian mixture models with diagonal covariance matrices. Given the case of n mixture components, and a k-dimensional feature vector x_t, as well as a diagonal covariance matrix, three parameters define the mixture components of state s_i:

- the weights $w^{(i)} = w_1^{(i)}, \dots, w_n^{(i)}$ for each mixture component
- a set of n k-dimensional vectors containing the means for each Gaussian $\mu_n^{(i)} = \mu_{n,1}^{(i)}, \dots, \mu_{n,k}^{(i)}$
- a set of n k-dimensional vectors containing the diagonal vector of the covariance matrix for each Gaussian $\sigma_n^{(i)} = \sigma_{n,1}^{(i)}, \dots, \sigma_{n,k}^{(i)}$

Consequently, the probability of state s_i emitting the observation x_t can be calculated as

$$p\left(x_t \mid s_i\right) = \sum_n w_n^{(i)} \mathcal{N}\left(x_t; \mu_n^{(i)}, \sigma_n^{(i)}\right). \tag{2.3}$$

Emission probabilities obtained by discriminative training, as for example provided by deep neural networks, normally provide state posterior probabilities $p\left(s_i \mid x_t\right)$ which are divided by $p\left(s_i\right)$ to obtain the emission probability $p\left(x_t \mid s_i\right)$, referred to as scaled likelihood estimation [BM94].

Properties of HMMs An HMM $\mathcal{H}$ makes three main assumptions about the process under consideration:

1. As the name suggests, a HMM models a Markov process which implies that HMMs satisfy the Markov property, i.e., the probability of being in state s_t at time t depends only on the state at $t-1$,

$$p\left(s_t \mid s_{t-1}, s_{t-2}, \dots, s_1\right) = p\left(s_t \mid s_{t-1}\right). \tag{2.4}$$

2. The transition matrix A is stationary, i.e., transition probabilities are independent of t.

3. The observations are *conditionally independent* from the preceding or subsequent history of the process.

$$p\left(X \mid s_1, s_2, \dots, s_t, \mathcal{H}\right) = \prod_t p\left(x_t \mid s_t, \mathcal{H}\right) \tag{2.5}$$

The three basic problems of HMMs There are three problems related to the use of HMMs: the evaluation problem, the decoding problem and the training problem ([Rab89]).

1. *Evaluation:* Given a HMM $\mathcal{H}$ and an observation $X = x_1, \dots, x_t$, calculate the probability $p\left(X \mid \mathcal{H}\right)$ that the given HMM produced this observation. This can be efficiently done by using the forward or forward-backward algorithm.

2. *Decoding:* Compute the most likely state sequence $q\left(t\right) = q_1, \dots, q_t$, with $q_t \in \{s_1, \dots, s_i\}$ for an observation $X = x_1, \dots, x_t$, given $\mathcal{H}$. Since the search space spanned by a HMM corresponds to a graph, this problem can be viewed as a

shortest-path problem and effectively solved by dynamic programming. $Q_t(i)$ keeps track of the probability of being in state i at time t, with $Q_{t+1}(j)$ corresponding to

$$Q_{t+1}(j) = p(x_{t+1} \mid s_j) \max_i Q_t(i)\, a_{ij}. \tag{2.6}$$

If for every time t and state s_j a bookkeeping list keeps pointers to the most likely preceding state s_j, the sequence $q(t)$ can be obtained by backtracking the most likely state of the last frame T from $q_T = \arg\max_i Q_T(i)$.

3. *Training:* Given a suitable initialization of the model parameters $\mathcal{H}$ of an HMM and a set of r training observations $\{X\}_r$, estimate the optimal set of parameters $\mathcal{H}$ according to an optimization criterion $F(\{X\}_r \mid \mathcal{H})$. The commonly used Maximum likelihood (ML) criterion aims at finding the parameters $\mathcal{H}$ that maximize the probability of the training sequences. ML estimation of the training data can be iteratively conducted by using the Expectation-maximization (EM) Algorithm.

Subword units and the lexicon For large vocabularies, it is impossible to learn specific acoustic models for each individual word. Each word w is therefore decomposed into subword units U, typically phones or context-dependent phones, for which large speech corpora provide a sufficient amount of training examples. This requires the *lexical model* as third source of knowledge, which contains the mapping from each word to a sequence of subword units. The lexicon also accounts for pronunciation variants of the same word.

The training problem in LVCSR Training parameters in LVCSR is challenging, since it requires labeled training data, normally corresponding to a huge collection of spoken utterances, providing an aligned subword sequence for each utterance. Since hand-labeling of speech data at the subword-level is a time consuming and difficult task, training of acoustic models in LVCSR usually relies on word-level transcriptions of the training utterances, that are decomposed into their corresponding subword sequences according to the lexicon and iteratively aligned to the training data. One common approach to initialize such a training scheme is to uniformly segment the training utterances into a sequence of subword units, further decomposed into its HMM states, while initializing the parameters of the continuous density HMMs by K-Means clustering. While ML estimation produces optimal generative models, it is desirable to train model parameters that minimize the word error rate (WER) of training utterances. While WER cannot be included in a differentiable objective function, many criteria have been proposed that allow discriminative training of the HMM parameters subject to an optimality criterion that approximates the WER, like minimum classification error (MCE), minimum phone error rate (MPE) or maximum mutual information (MMI) [SMMN01].

2.2.3. The language model

In Equation 2.1, $p(W)$ determines the prior probability of the word sequence $W = w_1, \ldots, w_n$. For large vocabulary ASR, $p(W)$ is usually provided by a probabilistic

n-gram language model. n-grams predict the probability of a word $p(w_k)$ at position k by taking the $n-1$ preceding words into account so that $p(W)$ is computed as

$$p(W) = \prod_k p(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}) . \tag{2.7}$$

There a three major issues concerning language modeling, which are only briefly mentioned in the following. The first issue is data-sparseness, i.e., a n-gram language model has to provide probabilities for all v^n possible combinations of a vocabulary of v words, while most of the n-grams are not present in the training data. Second, n-grams do not provide sufficient information about the long-distance dependencies that are present in speech grammars. And third, n-grams need to be regularized, since there will be a mismatch between the training and the test data.

2.2.4. Search in LVCSR

With the introduction of subword units U as the basic unit for acoustic modeling, which can be further decomposed into the states S of the corresponding HMM, Equation 2.1 can be rewritten as

$$p(X \mid W) = \sum_{\{S\}_W} p(X, S \mid W) . \tag{2.8}$$

$\{S\}_W$ corresponds to the set of all state sequences S that can form the word sequence W. $p(X \mid W)$ is commonly approximated as the most likely path via the *Viterbi approximation*

$$\hat{W} = \arg\max_W \left(p(W) \max_{\{S\}_W} p(X, S \mid W) \right) . \tag{2.9}$$

Thus, the search for the best word sequence $\hat{W}$ has been broken down into the search for a sequence of subword units $\hat{U}$, further decomposed into a state sequence $\hat{S}$.

Search strategies attempting to provide efficient solutions for Equation 2.9 can be roughly divided into static vs. dynamic expansion of the search network and time-synchronous vs. time-asynchronous decoding.

Static and dynamic expansion Without any optimization it is impossible to statically expand the search space for standard LVCSR problems due to hardware and time constraints. Thus, dynamic expansion of the search space using advanced pruning strategies to control the size of the expanded network have been the standard search strategies for several decades. With the more recent use of Weighted Finite State Transducers (WFST) [MPR02] for decoding, it is possible to combine all the knowledge sources from the lexicon, language model up to the HMM state level in one finite state automata using composition. By optimizing the graph prior to the decoding via determinization and minimization it is possible to use classical Viterbi decoding with minimized computational costs.

Dynamic expansion relies on either time-asynchronous or time-synchronous decoding. Time-asynchronous dynamic decoding, also referred to as stack decoding, keeps a stack of partial transcriptions of the utterance that are sequentially expanded. This corresponds to a *depth-first* strategy, since each word attached to the current hypothesis is expanded individually until a termination criterion is reached. Time-synchronous search expands all active state hypotheses in parallel and keeps simultaneously track of the individual word histories, following a *breadth-first* strategy.

Search space and decoding in this thesis The decoding strategy used in this thesis (see [NO99]) uses a time-synchronous decoding in connection with a pre-compiled prefix-tree lexicon. *Prefix-tree lexicons* are a compact representation of words which exploit the fact that many words share the same word stem. Words are organized into a phonetic-prefix tree and each arc of the tree splits words from the common beginning into individual successions until the terminal node is reached, which corresponds to the word ending.

Using a tree lexicon requires to keep separate copies of the lexicon in memory, with each copy corresponding to a different bi-gram word predecessor, given a tri-gram language model. The search space at each time frame then corresponds to a stack of several active prefix-trees, with each prefix tree pointing to a stack of active states. Thus, there are four dimensions to describe the full search space ([Aub02]): time, language-model state, phonetic arc and acoustic state. To limit the number of active hypotheses, several heuristic criteria prune away active hypotheses before entering a new frame to reduce the search space with minimal loss in performance. The pruning criteria used in this thesis are:

- Beam search: All active hypotheses with probabilities lower than a predefined fraction of the best current acoustic hypothesis are discarded.
- Histogram pruning: The number of surviving states is limited to a fixed number and the remaining states are pruned away.
- Language model look ahead: In connection with prefix-tree lexicons, the language probability of possible terminal nodes can be anticipated by looking ahead of the current state. Arcs that will result in a very low language probability can thus be pruned from the search space.

Multi-pass ASR The search for the best word hypothesis has to find the optimal compromise between the use of advanced statistical models and keeping the search space computationally tractable. For example, while the beam search algorithm used in this thesis uses very efficient pruning strategies, the decoding is limited to the use of intra-word triphones and tri-gram language models. A workaround that allows the usage of higher order statistical models is to use «simpler» models in a first decoding pass to create a set of n-best hypotheses, which can be converted into a word-graph, which is rescored in a second step [NS97].

It is evident that decoding errors from the first pass are propagated into the second pass and are unlikely to be recovered. The challenge is therefore to create word graphs that provide alternative hypotheses for parts with high acoustic ambiguity but minimal

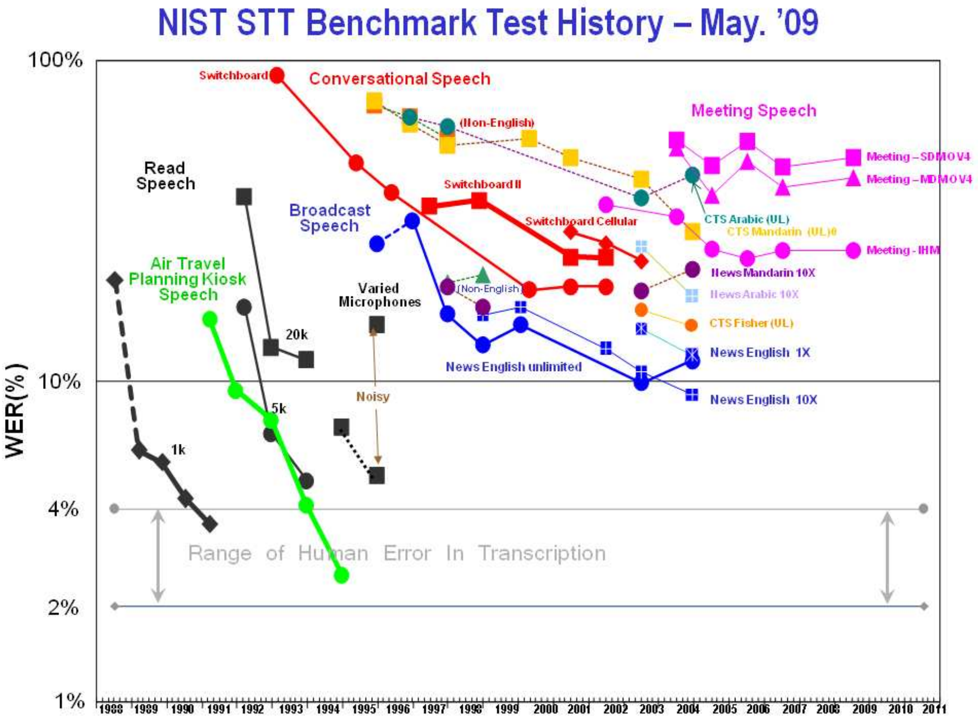


Figure 2.2.: NIST STT speech recognition benchmark tasks over the years (from [FAG08]).

density, to allow computational feasible decoding in the first and second pass without loosing relevant information.

2.3. Performance boundaries and issues of state-of-the-art ASR

Figure 2.2 displays the performance improvement of benchmark ASR systems over the years on different speech recognition tasks, as it has been recorded by the NIST STT Benchmark Tasks [FAG08]. It can be seen that while it was possible to reach human-like recognition performance for simple vocabularies in clean environments during the 90s, speech recognition for any large vocabulary tasks is still far from competing with human performance. The major bottleneck for further improvements in LVCSR, is its sensitivity to mismatch between training and testing utterances, which poses problems on several levels.

First, speech recognition is sensitive to different speaking styles, including changes in speaker, dialect, speaking rate or over-articulated speech. While huge databases allow to train precise models for different variants of speech, speech is an «open-set» problem [Lee04], i.e., it is not possible to provide training examples for all possible variations in speech. Another issue on the acoustic level is the variety of channels that can transform the speech signal and produce a mismatch between trained acoustic models and observed speech signal. While estimation of the noise, especially in the stationary case, can improve recognition, natural environments are still a huge obstacle towards robust speech recognition. On the level of the lexicon and language model, there are also mismatches between trained models and real utterances, due to the problem of data-sparseness.

2.4. Summary and conclusions

This chapter introduced modern statistical LVCSR consisting of a language model, which stores information about word sequences in form of n-grams, a lexicon that decomposes words into subword units and an acoustic model that maps acoustic information onto subword units. With HMMs as acoustic models, all knowledge sources can be compiled into a search graph to obtain a word hypothesis, by optimizing the probability of all knowledge sources at the same time.

Despite its success, several studies rightfully point out that statistical ASR might not be able to overcome the bottlenecks of automatic speech recognition evoked in the previous section. The main argument is that statistical ASR does not provide an accurate model of the speech process and performance relies heavily on training models on the same type of speech that is to be recognized. Indeed, it is unlikely that increasing the training data can account for all possible sources of speech variability. Additionally, statistical speech recognition performes poorly for under-resourced languages which can not provide sufficient acoustic or linguistic data to use many of the tools that are common in ASR frameworks. Since there is an obvious gap between human speech processing and the acoustic modeling in state-of-the-art ASR and humans can effortlessly overcome most of the problems that ASR fails to solve, it is straightforward to assume that using more phonetic knowledge in acoustic modeling might push automatic speech recognition towards human-like performance on difficult recognition tasks.

The following chapter will introduce the topic of this thesis by first summarizing the limitations of standard HMM-based acoustic models in capturing the complete spectrum of phonetic relevant information, before reviewing existing phonetically motivated alternative modeling approaches and introducing landmarks as promising concept for building a bridge between statistical and knowledge-based ASR.

3. Towards landmark-driven ASR: Motivation and related work

The first purpose of this chapter is to give a detailed overview of the most frequently cited shortcomings of HMM-based acoustic modeling with regard to its ability to capture relevant phonetic information, but also of the advantages of HMMs and statistical ASR, which made HMMs the dominant acoustic modeling technique in the first place. This overview is followed by a review of related work that also discusses the motivation behind integrating phonetic knowledge into HMM-based ASR, rather than replacing HMMs with a new acoustic model.

The second purpose is to introduce the landmark-based acoustic modeling paradigm, which will be the central subject of this thesis. Landmark-based ASR is based on an alternative speech representation, representing speech as a sequence of discrete time events that are associated with phonetic labels. While it will be seen that landmarks are not likely to replace HMMs as acoustic models in the near future, phonetic landmarks possess many desired properties that make them a promising medium that could possibly integrate phonetic knowledge into the search for the best word hypothesis in state-of-the-art ASR.

3.1. Advantages and disadvantages of HMM-based acoustic modeling

HMMs impose multiple constraints onto the speech process that violate basic principles of human speech production and perception (see chapters 2 and 1), yet HMMs inside statistical ASR are the state-of-the-art approach to acoustic modeling. Therefore, the following section first reviews the advantages of HMMs and the statistical modeling framework, before pointing out the most common points of criticism.

3.1.1. Advantages

Phonetic knowledge in statistical ASR The architecture of statistical ASR and its «ecosystem» [SC12] of signal processing and adaption methods account for several basic, nonetheless important properties of speech. The acoustic observations, for example MFCC vectors, are adapted according to the sensibility of the human ear and there are several methods to reduce the influence of the channel and speaker onto the acoustic observations. HMMs as the basic acoustic model account for the temporal variability of speech sounds, by modeling speech units as time-variable segments and models can equally be adapted towards speakers and channels. The decomposition of words into subword units acknowledges speech being composed of basic building blocks and allows different pronunciation variants of words. Subword units like triphones can take the context dependency of speech units into account and training triphone-based models additionally relies on phonetic knowledge by selecting phonetic questions to determine shared states of triphones [YOW94].

Advantages of the architecture The basic architecture of statistical ASR can be decomposed into several modular tasks, from signal processing, acoustic modeling, language modeling to decoding and rescoring. This has allowed the ongoing improvement of specialized tasks inside the ASR framework that can be seamlessly integrated into existing systems.

Lexical, language and acoustic model form a composite Markov chain during the search process that incorporates all levels of knowledge in one big network. This allows parallel usage of all knowledge sources during the search for the best word hypothesis and avoids a hierarchical process that segments the speech signal sequentially on the acoustic, semantic and syntactic level, since the decoding does implicitly provide the segmentation of speech into its states, subword units and words. This makes the search process robust against inaccuracies at each level and acoustically variable speech might still be correctly recognized due to accurate language models.

Effective training framework While all models in statistical ASR, i.e., acoustic, language and lexical models, are very simple models that make unrealistic assumptions about the complex speech process, they do not rely on the sometimes limited or incomplete corpus of phonetic knowledge, but can obtain the necessary parameters by optimizing the models according to the training data [HH94].

Since effective training procedures like the Baum-Welch re-estimation of the HMM parameters do only rely on word-level transcriptions of training utterances, modern systems can estimate statistical parameters from huge amounts of training data, without needing hand-labeled phonetic or citation-phonemic speech transcriptions.

3.1.2. Disadvantages

The speech model of statistical ASR is often referred to as the «beads-on-a-string» model of speech, i.e., speech is modeled as a sequence of fixed-size time frames, with each frame being attached with a phone-based speech label and reading the temporal sequence of speech labels results in the citation-phonemic pronunciation of word sequences (see [Ost99]). This simple model of the speech process has clear limitations when it comes to capturing relevant phonetic information and incorporating basic phonetic knowledge (see Chapter 1). The main points of criticism concerning the acoustic observations and symbolic representation used in this model, as well as the limitations of the HMM modeling paradigm are pointed out in the following paragraphs.

The homogeneous observation space Speech theories agree on the fact that the human speech perception is based on multiple heterogeneous information streams. While the exact mechanisms are not known, it is assumed that the temporal and spectral variability of the processing streams can account for the human ability to understand speech in adverse conditions and robust speech understanding would not be possible if only a single stream of acoustic information would be used for processing the acoustic signal [CLA05]. Indeed, experiments show that it is possible to significantly alter the spectral information of a speech signal without diminishing the human capability to perceive the intended utterance, as long as certain acoustic cues are preserved in the signal [Coo06, Lip97]. Relying on a single stream of frame-based spectral information in modern ASR might be one of the reasons that ASR is still sensitive to all kinds of spectral distortions.

Additionally, modern ASR does not account for the hierarchical processing of these heterogeneous information streams, which is supposed to convert the raw audio signal into intermediate representations before higher level linguistic meaning is derived. Some studies (e.g., [MHP12]) argue that the recent leap in LVCSR performance by using deep neural networks to predict the state-based emission probabilities can be partly due to the deep architecture of the network that allows to capture higher order structure of the acoustic signal.

Another drawback resulting from the acoustic observations used inside the beads-on-a-string model of speech is the fact that the acoustic likelihood of the word hypothesis is derived by summing up the classification scores of all individual frame-likelihoods. By continuously judging the probability of speech labels according to short-term spectral observations, the most likely speech hypothesis will be influenced by spectral observations that are noisy, perceptually irrelevant and not necessarily indicative of the presence of any higher level speech unit.

Articulatory information and speech representation While context-dependent phone models account for coarticulation, they only describe the effects of coarticulation, rather than modeling the process itself (see [FWK07, Ost99]). Thus, many studies argue that direct modeling of the speech articulators or adding an articulatory feature representation to the regular phone-based speech representation can account for many speech phenomena that are not captured by standard phone-based models, especially coarticulation and pronunciation variants of the same word.

Properties of HMMs The topology of HMMs imposes several constraints on the speech process that violate basic phonetic principles. The most obvious violation is the assumption of the emission probability being only dependent on the frame-based acoustic observation and the state. In fact, the acoustic observations are dependent on the physical constraints of the vocal tract and the slowly varying articulatory movements. Therefore, subsequent acoustic observation vectors are highly correlated. In contrast to that, HMMs imply that the emission probability only depends on the piecewise stationary acoustic observation and does consider each observation as drawn from a probability distribution estimated on a plurality of speakers.

Since the transition probabilities are time-invariant probabilities, HMMs implicitly model the duration of subword units as a geometrical distribution, i.e., the duration probability decreases exponentially with the duration of the subword unit, which is clearly not an appropriate temporal model for speech units.

3.2. Phonetically motivated acoustic modeling

With regard to the discussed limitations of HMMs, several research projects aimed at replacing or extending HMM-based acoustic modeling by phonetically motivated models. This section divides these approaches into approaches that extend the HMM-based modeling paradigm to account for phonetic information inside new acoustic models and approaches that combine regular HMM-based acoustic models with additional phonetic information.

3.2.1. Extending HMMs

Given the success of HMMs as acoustic models, it seems natural that many studies attempt to extend the existing architecture of HMMs or use related models that can account for desirable phonetic properties, but embed these models into a similar statistical decoding framework. The following section gives a closer look onto approaches that attempt to integrate articulatory information into statistical ASR, replace HMMs by the more the flexible dynamic Bayesian networks and use segment-based acoustic models.

Articulatory information in HMMs While direct measurements of articulatory movements using tomography, ultrasound or similar methods is an interesting method to gain detailed knowledge of speech production, data acquisition is too complex to generate sufficient data for LVCSR subword models [KFL$^+$07]. Thus, articulatory information inside ASR is usually based on imposing articulatory speech models on available speech corpora and estimating the articulatory parameters by articulatory-acoustic inversion [SK86, ACMT78]. [Wak79] proposed a method for estimating vocal-tract shape parameters, which was extended in [Krs00, Krs99] for inverting the acoustic signal into a parametrized vocal tract representation, which was used to train HMMs for medium vocabulary ASR, which did not improve speech recognition compared to the baseline.

Trajectory HMMs replace the rather heuristic first and second order derivatives of the acoustic observation vector that provide temporal context information in most ASR systems by defining dynamic features as a function of the static observation vector. Using trajectory HMMs to model these production related parameters was used in [ZTK07] together with modified Viterbi decoding for rescoring the hypotheses obtained from standard HMM-based first pass decoding which improved phoneme accuracy of a phoneme recognition task.

The most common usage of articulatory information is the use of distinctive feature transcriptions obtained from canonical mapping of phone labels as an alternative to standard phone-based speech representations. Using distinctive features inside HMMs was studied in [KFS02] where parallel independent statistical classifiers for distinctive features are trained and combined to a state-level score inside standard subword based

LVCSR. The results indicate improvements over standard phone-based models at a low signal-to-noise ratio.

Dynamic Bayesian networks Dynamic Bayesian networks (DBNs) have been used to extend the dependencies of classical HMMs to integrate articulatory knowledge directly in the architecture of the acoustic model. A dynamic Bayesian network is an acyclic graph, with nodes corresponding to random variables. An arc represents a conditional relationship between variables, which are visible or hidden and continuous or discrete. An HMM is similar to a very simple DBN, yet DBNs are capable of modeling arbitrary conditional dependencies between observed and hidden variables at each time frame. Acoustic modeling using DBNs can account for dependencies on speaking style or rate ([ZR98]), but also phonetically motivated dependencies which led to a variety of DBN architectures that include distinctive feature models in acoustic DBNs.

[LGB03] proposes a flexible DBN that adds exponentially weighted distinctive feature scores to standard phone-based emission probabilities. If the DBN uses unfactored distinctive feature models, the modified decoding gives a slight improvement over the HMM-baseline, especially in the presence of noise. While modeling the dependencies between distinctive features increases feature recognition accuracy, compared to independent HMM-based feature modeling, ANNs still outperform DBNs on feature recognition accuracy in [FWK04]. [FWK07] presents a hybrid ANN/DBN architecture that improves feature recognition accuracy compared to an ANN/HMM baseline. Additionally, this study provides a method to train asynchronous distinctive features in a data-driven way, which avoids closely following the beads-on-a-string modeling paradigm.

Segment-based ASR Segment-based ASR proposes the use of acoustic information extracted from temporal variable segments rather than fixed short-term time frames. The main motivation for using segments instead of frame-based acoustic observations is that segments take into account the statistical dependence of short-term spectral observations, resulting in an acoustic model that is much closer to the process of speech production. Segment-based ASR relaxes the independence assumptions of successive spectral observations and segments are supposed to provide a more accurate temporal model for subword units, changing the geometric temporal modeling of HMMs into an explicit parametric or non-parametric distribution [Ost99].

In [Ost99], acoustic observations are jointly modeled by a duration model and a segment model that provides probabilities for temporal variable acoustic observations, with the acoustic observations being obtained by warping frame-based acoustic information on regions of feature trajectories.

The SUMMIT speech recognition system [Gla03] is also a segment based speech recognizer, providing a probabilistic framework that relaxes the frame-based conditional independence into a more flexible segment-based conditional independence assumption and allows the use of heterogeneous segment-based acoustic observations. In this framework, the acoustic observations are organized as multi-level speech segments, ranging from short spectral variations to long and spectrally constant segments. This approach improves phone recognition accuracy compared to conventional HMMs.

3.2.2. Integrating phonetic knowledge into HMM-based ASR

Integration of phonetic knowledge into standard HMM-based ASR has been done on every level of the statistical speech recognition framework, from feature-level over acoustic model combination to rescoring approaches.

Feature fusion approaches using articulatory features Feature fusion is used for the integration of binary articulatory features into HMM-based ASR in [Eid01], by predicting articulatory features for every frame in the speech signal. To obtain a modified acoustic observation vector, the log-likelihood ratios of the articulatory feature classifiers are concatenated with the original MFCCs to retrain articulatory feature models. This procedure is repeated once more to obtain the final modified observation vector that is used to train the acoustic models. This approach improved the performance of a small vocabulary recognition task in noisy car environments.

Combination of articulatory features and emission probabilities Several studies [SMSW03, Met05, Met06, MW02] use a weighted linear combination approach that combines the emission probabilities of the ASR system with articulatory feature GMMs that model frame-based articulatory features with two GMMs, corresponding to the absence and presence of the feature.

While this combination approach generally improves speech recognition compared to the baseline system, there is few sensitivity towards different feature selection strategies [MW02] and towards the different methods for estimating the stream weights, including empirically fixed weights and weights obtained from optimizing different discriminative training criteria [Met05, SMSW03]. Yet, weights obtained from discriminative training perform better than empirical weights when cross lingual features are adapted to monolingual acoustic models [SMSW03].

Automatic speech attribute transcription The automatic speech attribute transcription (ASAT) project aims at extending statistical ASR from «knowledge-ignorant» to «knowledge-rich» acoustic modeling, by integrating frame-based feature detectors into statistical ASR (see [Lee04]). Several studies have been published in the context of this project, which propose a variety of acoustic observation vectors and several frame-based distinctive feature detectors, using HMMs ([SL09]), ANNs, SVMs or MLPs ([BQH$^+$07]) that provide log-likelihood ratios for distinctive features. The continuous stream of frame-based likelihood scores is used to predict phone posteriors using conditional random fields (CRFs) or ANNs, that are used for lattice rescoring.

Despite some advantages and small increases in recognition performance, no acoustic model or framework that was presented in this section has been able to impose itself as the new standard in acoustic modeling. Indeed, all presented approaches come with several drawbacks, which make them less appealing compared to HMMs and the increase in the amount of phonetic knowledge incorporated in these models is usually quite limited, which is discussed in the final paragraph of this section.

3.2.3. Discussion

Many presented approaches showed that accounting for phonetic knowledge inside new acoustic models usually requires more complex training and decoding methods. For example, the large degree of freedom of DBNs concerning modeling different dependencies can become very complex and training these models encounters the problem of data-sparseness so that training and decoding rapidly become prohibitive. Segment-based ASR also increases the complexity of the search for the best word hypothesis, since one has to take into account all possible segmentation sequences during decoding. Additionally, temporal information by explicit duration models seems to be highly variable and thus does not propagate into improved word hypotheses. Articulatory motivated models and DBNs keep the standard frame-based speech model, but change the labels from phone-based into articulatory or distinctive feature representations. This similarity in the speech model might explain the relatively limited gain in performance of those methods compared to HMM baseline systems.

Most approaches that attempt to integrate phonetic models into HMM-based ASR also use frame-based feature representations and corresponding classifiers in addition to standard phone-based models. While these approaches seem to provide a constant gain over HMM-based ASR, the modeling paradigm remains unchanged, with the exception of an additional layer of feature labels. Clearly, this modeling approach propagates most of the disadvantages of the acoustic modeling paradigm of HMMs into the phonetic models, so there is the legitimate question whether phonetic knowledge might need different modeling techniques to provide complementary information to the standard ASR models.

3.3. Landmark-based approaches to ASR

While all previously discussed approaches rely on modeling the signal as a sequence of continuous frames that can be concatenated to citation-phonemic speech transcriptions, landmark-based approaches model speech as a sequence of discrete events, referred to as landmarks, that follow each other at irregular time intervals. Each landmark indicates the presence of a relevant phonetic event, for example an acoustic cue, that is associated with one or a bundle of phonetic labels. Higher level speech units can be derived by decoding the resulting sequence of phonetic labels. Landmark detection algorithms can generally be divided into two parts, with the first part corresponding to the detection of perceptually relevant time instances in the acoustic signal and the second part consisting in evaluating the acoustic content in the vicinity of these landmarks to attach the landmark with one or more phonetic speech labels.

The following summary discusses landmark-based approaches to ASR that are based on expert rules, as well as approaches that use landmarks inside statistical frameworks.

Lexical access from features Lexical access from features (LAFF) [Ste02] is a strictly rule-based landmark detection framework, which can be categorized into acoustic-phonetic approaches to speech recognition, which attempt to make use of phonetic knowledge by detecting and classifying speech units according to expert rules. In LAFF, expert knowledge is used to derive a number of acoustic correlates for the manner of articulation gestures, for example maxima in intensity or spectral changes, which are used to detect perceptually relevant landmarks in the signal. Several acoustic cues are measured to reveal additional information about the full set of distinctive features that are used to describe each landmark. LAFF has not been applied to speech recognition tasks, since expert rules cannot cope with the variability of speech. Moreover, rule-based systems lack the possibility to optimize their models using mathematical optimization [Jel97].

Event-based speech recognition In [JEW08], landmark-based ASR is extended towards a statistical framework. The proposed algorithm consists of a first classification and segmentation step, that uses frame-based SVM classifiers in connection with a probabilistic Viterbi segmentation to obtain a sequence of syllabic, sonorant and continuant segments. Specific knowledge-based acoustic parameters are then used to determine the exact landmark positions for syllabic peaks, vowel onsets, bursts, syllabic dips, onsets and offsets of sonorant consonants as well as fricative onsets and offsets. The obtained sequence of landmarks is used to calculate the likelihood of isolated words and their canonical landmark representation in connection with broad phonetic segmentation likelihoods and duration probabilities.

Landmarks in SUMMIT While the SUMMIT speech recognition system is a segment-based speech recognizer (see Section 3.2.1), it provides optionally the possibility to integrate landmarks into the probabilistic decoding. Landmarks are essentially additional probabilistic observations, either at segment boundaries or inside speech segments, emphasizing the acoustic information at these presumably perceptually important parts of the speech signal. Speech recognition experiments show that landmarks provide additional information during decoding and help to improve the overall recognition performance.

Landmarks for word graph rescoring The 2004 John Hopkins Summer Workshop [HJBB$^+$05] studied the use of phonetic landmarks for rescoring word graphs obtained from standard HMM-based ASR. Landmarks correspond to the positive output of binary frame-based classifiers and indicate onsets and offsets of phonetic classes like fricatives, sonorants and vowels as well as several phonetic nuclei. The landmarks were employed in three different settings. First, landmarks were used together with a generative feature-based pronunciation model by combining pronunciation models with SVM classifier outputs. The second approach was discriminative rescoring of lattices by using the obtained landmarks to rescore confusion networks of a first-pass baseline decoder. The third approach equally used landmarks for lattice rescoring, but integrates landmarks with the language and acoustic models by weighted combination. While all proposed approaches were able to correct some word errors, they produced an equal amount of new word errors, so that there was no overall improvement.

Point process models of speech Landmarks have also been exploited for keyword spotting in the context of point process modeling of speech [JN08, JN09a, JN09b, NJRT12]. Point process models correspond to landmark detectors that mimic the firing patterns of neurons, which are combined by statistical models to predict phonetic sequences. In [JN08], landmark detection functions derived from homogeneous frame-based predictions have been integrated in a hierarchical framework which first segments speech into sonorant and obstruent regions before these segments are decoded with a probabilistic segment decoding algorithm to obtain a broad phonetic class transcription of the speech utterance. In [JN09a], phone-based landmark profiles are jointly integrated using statistical models to decode obstruent speech regions.

Phonetic class detectors Additionally to the presented landmark-based frameworks, which are all linked to a specific application, several studies focus on the detection of phonetic classes without immediate benefit for speech applications.

One of the most studied phonetic classes are plosives, because of their distinctive acoustic cues which have been exploited in several detection algorithms. [ZHJB04] combines formant estimation by several algorithms to detect different types of stop consonants. [KCB01] uses a detection and classification framework that uses several hierarchical decisions to reduce insertions. [LW11] uses the two-dimensional cepstrum to capture the dynamics of plosives in connection with random forest classifiers. Other notable phonetic classes studied involve the detection of vowels, which [How00] identifies by examining the maxima of first formant energy bands, nasality, which equally relies on energy measurements in [PEW04] and fricatives, which are detected in [RDL10] by evaluating a distance measure of a cepstrogram based template-matching method.

Discussion While the presented approaches to landmark detection and landmark-based ASR offer interesting new perspectives on speech recognition and speech processing, they are unlikely to compete with HMM-based acoustic modeling in the short term, for several reasons. First, if landmark detection makes strong use of phonetically motivated models, they usually can only model broad phonetic speech classes, for which acoustic cues are sufficiently studied, which prevents the use of these systems for sophisticated recognition tasks. Furthermore, landmark detection front-ends are likely to miss speech events during detection, which is difficult to overcome in later stages. The last issue is the lack of a decoding framework that combines the language model with landmarks to obtain reliable word hypotheses. It is straightforward to conclude that landmarks might be more beneficial for ASR, when they are integrated into the standard statistical ASR framework, which will be the subject of this thesis and is motivated in the concluding section of this chapter.

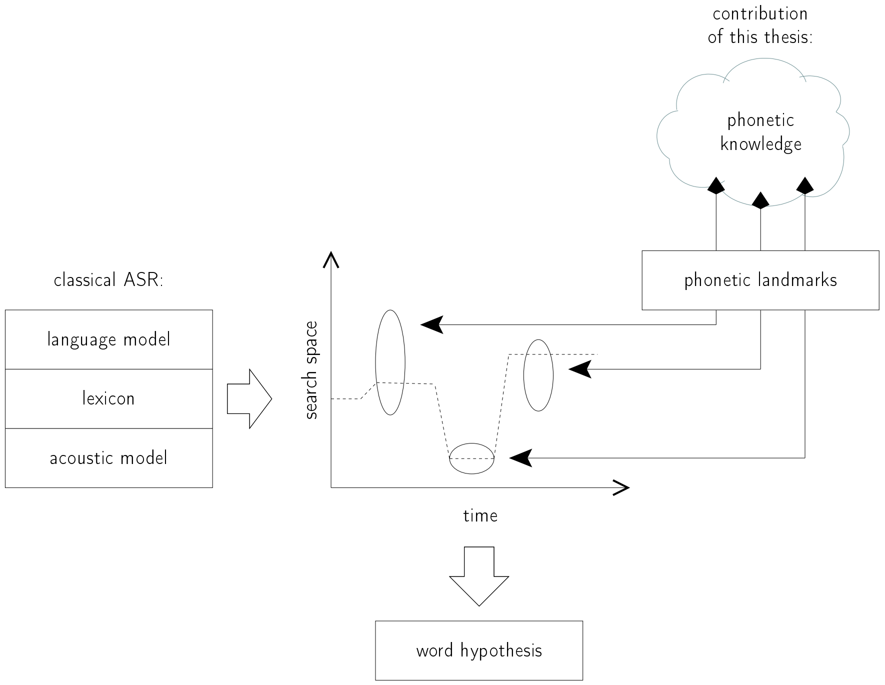


Figure 3.1.: The abstract concept of landmark-driven ASR as it is used in this thesis. The two main components are the baseline ASR system, as it has been presented in Chapter 2 and the contribution of this thesis consists in adding a landmark detection component that can modify the search for the best word hypothesis of the baseline system by biasing the search towards detected phonetic information.

3.4. Objectives of this thesis

The objective of this thesis is to study the integration of phonetic landmark sequences, as they are used in landmark-based ASR, into the standard Viterbi decoding of statistical ASR. The goal is to use the information provided by landmarks in order to obtain better word graphs and, potentially, a faster decoding time. The main hypothesis that motivates the use of landmarks in combination with statistical ASR in this thesis is that landmarks can rely on phonetically motivated models that are able to model *some phonetic classes* at *certain time instances* more accurately than it can be done by the emission probabilities of standard HMM. Indeed, the results from related work often show that landmarks provide good detection results for broad phonetic classes, where heuristics rely on relatively few parameters and acoustic cues and their effect on perception is well studied. While missing speech events due to speech variability might have detrimental effects on the speech recognition performance if landmarks are the sole paradigm for acoustic modeling, hybrid landmark-driven HMM-based ASR systems might rely on its statistical power at frames where landmark information is missing, but take into account the provided phonetic information if landmarks are present in the signal. Thus, landmarks can convert a potential weakness of the body of phonetic work available into a strength: Use the incomplete body of phonetic knowledge if possible, but leave the work to the statistical ASR framework when there is too much acoustic ambiguity in the speech signal.

While many related studies focus on the use of phonetic knowledge or landmarks for rescoring word graphs, this thesis aims at studying the integration of landmarks into the search space of the first pass of statistical ASR. The landmark-driven speech recognition concept that will be pursued in the remainder of this thesis is displayed in Figure 3.1, with landmarks as a third model, besides acoustic and language model. In this concept, landmark detection font-ends provide information about the presence or absence of phonetic classes at certain time frames, which can be used to bias the search towards including or excluding some states of the search space.

There are two main problems that are in the focus of this thesis. First, the thesis examines the practical issues in designing landmark detection front-ends that are able to incorporate phonetic knowledge into their models and proposes several landmark detection approaches. The second problem that will be discussed is how to use the output of landmark detection front-ends to modify the decoding step of standard ASR and how the choice of the integration method influences the amount of phonetic knowledge that can actually be included in phonetic detection front-ends.

4. A general landmark detection framework for landmark-driven ASR

The first purpose of this chapter is to introduce the general landmark detection framework that will be used for the remainder of this thesis, consisting of a bank of phonetic detection functions that correlate with the presence of phonetic labels. These detection functions are further processed to a set of binary landmarks that are integrated into the search for the best word hypothesis inside a standard HMM-based ASR system. The focus will be on comparing two different landmark detection architectures, with one architecture consisting of an individual detection front-end for each phonetic class and the other using one shared front-end for all phonetic classes. It will be seen that individual front-ends for each phonetic class allow considerable freedom concerning designing individual landmark detectors according to phonetic knowledge, but using one shared front-end for all phonetic classes might be easier to integrate into statistical ASR. The second purpose of this chapter is to introduce the integration methods that will be used in this thesis to integrate the detected landmarks into the search for the best word hypothesis.

4.1. General landmark detection framework

The basic building blocks of the general landmark detection framework of this thesis are displayed in Figure 4.1. The main part of this framework are k landmark detection functions $x_k(\tau_k)$, also referred to as k knowledge sources, which are only defined at selected time instances $\tau_k = \{t_1, \dots, t_{n_k}\}$. $x_k(\tau_k)$ indicates the presence of a phonetic label c_k by a score $x_k(\tau_k) \in]-\infty, +\infty[$, assuming a positive correlation of $x_k(\tau_k)$ and c_k. A priori, there is no limit in the number of knowledge sources, but it is straightforward to assume that at least one knowledge source that is present with $k \geq 1$. The phonetic label c_k always corresponds to a natural speech class or more generally to an ensemble of phones or non-speech symbols $\mathcal{S}_k$, so that with $\mathcal{P}$ corresponding to the context-independent phone inventory of a phone-based speech recognizer, a knowledge source k corresponds to a subset of phones $\mathcal{S}_k \subset \mathcal{P}$.

To filter out unreliable time frames $t \in \tau_k$, usually corresponding to low values of $x_k(\tau_k)$, the k knowledge sources are converted into k binary landmark indicator functions $\Lambda_k(t) \in \{0, 1\}$, by either jointly or individually processing each detection function $x_k(\tau_k)$. The obtained landmarks $\Lambda_k(t)$ are then used to bias the search of a phone-based ASR system towards incorporating the phones $\mathcal{S}_k$ in the word hypotheses at t when $\Lambda_k(t) = 1$ (see Figure 4.2 for an abstract example).

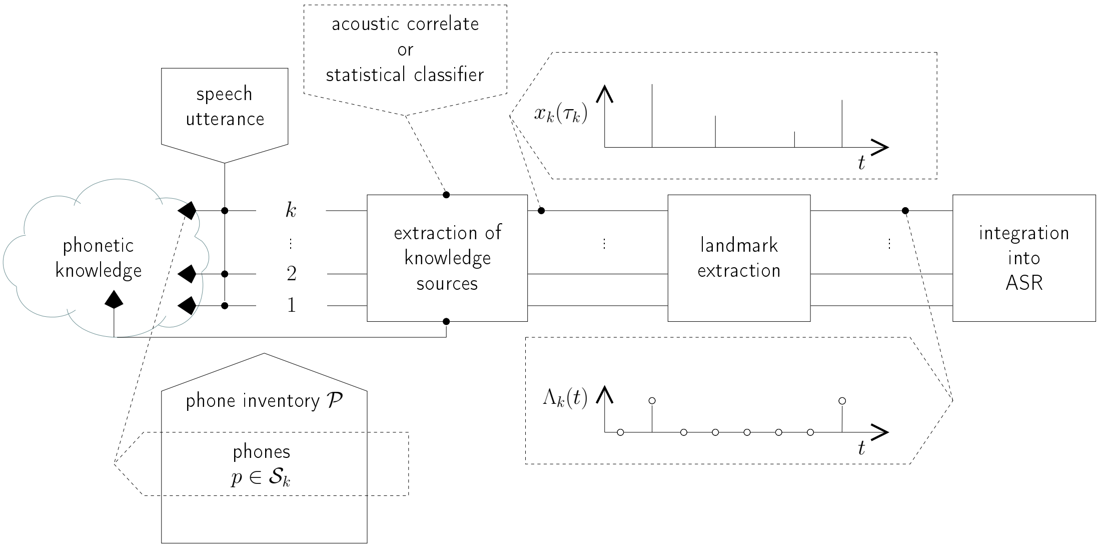


Figure 4.1.: The general landmark detection framework of this thesis, consisting of k landmark-based phonetic detection functions $x_k(\tau_k)$ which indicate the presence of a natural speech class corresponding to a set of phones $\mathcal{S}_k$. A priori, each detection function can be the result of an acoustic correlate or a statistical classifier that models either each phonetic class individually or all k classes together. The k detection functions are converted into a set of k binary landmark functions $\Lambda_k(t)$ that indicate the presence of $\mathcal{S}_k$ at t by $\Lambda_k(t) = 1$.

Continuous and sporadic detection functions All k detection functions $x_k(\tau_k)$ follow the landmark-based modeling paradigm, since they are only defined for n_k frames with $\tau_k = \{t_1, \dots, t_{n_k}\}$ and $n_k \leq T$. To distinguish these detection functions from continuous frame-based detection functions $x_k(t)$, this thesis will utilize the term *sporadic* knowledge for $x_k(\tau_k)$. A priori, the details on how each detection function determines landmark positions τ_k are individually different from detection front-end to detection front-end and might range from post-processing of continuous detection functions to more sophisticated detection methods. While the main focus of this thesis is the development of detection and integration frameworks for sporadic detection functions, continuous detection functions $x_k(t)$ will additionally be used to benchmark landmark-based detection functions $x_k(\tau_k)$.

Phonetic labels In this thesis, phonetic labels c_k do always correspond to natural speech classes, which can always be reduced to a set of phones $\mathcal{S}_k$ of the phone alphabet of the speech recognizer. The reason for limiting oneself to an ensemble of phones are twofold. First, detection functions usually rely on statistical classifiers that are trained on force aligned phone models that are converted into phonetic representations. Second, the detected phonetic labels have also to be integrated into the phone-based search space

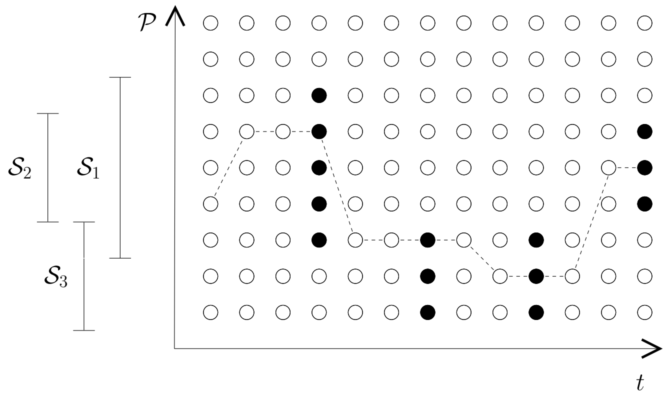


Figure 4.2.: Simplified example of biasing the search for the best path (displayed as a dotted line) according to three phonetic classes that correspond to the set of phones $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$. In this example the union of $\bigcup_k \mathcal{S}_k$ does not cover the full set $\mathcal{P}$.

of the ASR system. While these practical reasons are without any real alternative, there are several points worth mentioning in this context.

First, relying on the speech labels obtained from the beads-on-a-string speech model of statistical ASR naturally also means to adapt its drawbacks and inaccuracies as they have been discussed in Chapter 1 and Chapter 3. Additionally, this definition excludes all labels that could be associated with a transition between phonetic classes and it should be noted that $x_k(\tau_k)$ always indicates the nucleus of a phonetic class. This is different from related work (e.g., [JEW08]), where detecting and post-processing onset, nucleus and offset of a phonetic class was part of phonetic landmark detection. Another noteworthy constraint of the defined phonetic knowledge is the temporal resolution, which has to be synchronized with the temporal resolution of the frames of the ASR system. It also should be mentioned that this thesis only studies the integration of phonetic landmarks in the context of classical phone-based ASR systems. Other subword units, for example syllable-based models (e.g., [GHP$^+$01, JDM97]) are not studied in this thesis.

The next section discusses how to obtain $x_k(\tau_k)$ and two basic architectures of detection front-ends.

4.2. Knowledge sources in this thesis

This section explains how detection functions $x_k(\tau)$ can be obtained and presents two fundamentally different landmark detection frameworks, with one framework relying on a collection of individually designed landmark detectors and the other framework using one single detector that outputs multiple detection functions.

4.2.1. Calculating detection functions

In related work, $x_k(\tau_k)$ can either correspond to the output of a statistical classifier or an acoustic correlate.

Acoustic correlates Acoustic correlates correspond to acoustic measurements that are optionally processed with simple arithmetics to provide a score that correlates with the presence of a certain phonetic class. Acoustic correlates, for example the use of high-frequency energy as a correlate for fricatives, are usually employed as acoustic observations inside statistical classifiers. Nevertheless, every external system that provides a measurement correlating with the presence of a phonetic class, for example a distance measure resulting from a pattern matching algorithm (e.g., [RDL10]), can also be interpreted as an acoustic correlate.

Statistical classifiers Statistical classifiers are methods that rely on supervised training of statistical models using acoustic observations of training examples to provide an estimation for the presence or absence of $\mathcal{S}_k$. Classifiers can be divided into generative and discriminative classifiers. Generative classifiers learn the full joint distribution of a class label c and the associated attributes $\mathbf{x}$ and thus are able to predict an unknown instance given the learned distribution by applying Bayesian decision theory. In contrast to that, discriminative classifiers learn to separate the features space of two or multiple classes by minimizing a cost function during training. In that case, probability estimates or prediction scores $p(c \mid \mathbf{x})$ have to be derived by interpreting the output of the classifier, for example by normalizing the output layer of a multi-layer perceptron or estimating the distance to the decision boundary in the case of support vector machines.

Statistical classifiers for landmark detection are usually embedded in front-ends of different complexity that incorporate phonetic knowledge in their design, as it has been already shown in Section 3.3. Examples for possible design choices correspond to converting the speech waveform into a representation that allows to extract meaningful acoustic observations or to detect acoustic cues and methods to post-process noisy detection functions.

4.2.2. Individual and shared detection front-ends

The definition of $x_k(\tau_k)$ as it has been proposed in Section 4.1 does not set many constraints on potential detection functions and does account for most front-ends that have been proposed in the literature to produce single outputs $x_k(\tau_k)$ or multiple outputs $\{x_k(\tau_k)\}_k$. Of course, this thesis cannot explore all possible combinations and systems that can produce k detection functions. Therefore two different architectures that lie on opposite spectra of potential front-ends that provide $\{x_k(\tau_k)\}_k$ are of specific interest for the remainder of this thesis.

- The first architecture consists of an ensemble of k individual front-ends that produce independently k detection functions $\{x_k(\tau_k)\}_k$. Each system relies on individual acoustic representations and detection algorithms, specialized in detecting an arbitrary phonetic class k and there is no information exchange between different detection functions until integrating them into the search for the best word hypothesis. This heterogeneous processing results in asynchronous landmarks $\tau_1 \neq \tau_2 \neq \ldots \neq \tau_k$ and scores $x_k(t)$ that are not directly comparable. An example for such individual modeling of phonetic classes resulting in asynchronous landmarks would be detecting vowel landmarks by evaluating formant information, which would restrict landmarks to appear in voiced speech and detecting plosive landmarks by detection and evaluation of the voiced-onset time.

- The second architecture consists of one single front-end that follows one common modeling technique to produce multiple sporadic detection functions, like it is for example done in [JN09a, MH04]. Landmark detection based on a single front-end, as implemented in this thesis, can be simplified to two steps. In a first step, potential landmark candidates $\tau = t_1, \ldots, t_n$ are determined, before each landmark is associated with a score $x_k(\tau)$ for k phonetic classes, obtained by a multi-class classifier. The k classes are non-overlapping with $\bigcap_k \mathcal{S}_k = \emptyset$ and do cover the whole phone inventory $\bigcup_k \mathcal{S}_k = \mathcal{P}$. The scores $x_k(t)$ at each landmark $t \in \tau$, while not necessarily strict probabilities, can rank the k classes at $t \in \tau$ according to the confidence that the phone of the correct path $q_{\mathcal{P}}(t)$ at frame t is included in $\mathcal{S}_k$ with $q_{\mathcal{P}}(t) \in \mathcal{S}_k$[1].

While the second framework follows the landmark-modeling paradigm by avoiding to describe frames with high acoustic ambiguity and changing the prediction to phonetic classes instead of phones, the first framework clearly allows more freedom in adapting each individual front-end towards each phonetic class k under consideration. While possible methods for integrating k detection functions $\{x_k(\tau_k)\}_k$ are discussed in detail in Section 4.4, ideally both methods should enable accurate probability estimates about the correct path at frame t. These probability estimates should for example allow to compare the two hypothesis $H_1 : q_{\mathcal{P}}(t) \in \mathcal{S}_k$ and $H_2 : q_{\mathcal{P}}(t) \notin \mathcal{S}_k$ at a frame $t \in \tau_k$ or map $\{x_k(t)\}_k$ at frame t directly onto phone posteriors, to modify the search for the best word hypothesis accordingly. While the output $\{x_k(\tau)\}_k$ of the shared front-end can be directly interpreted as probability estimates for the phones $p \in \mathcal{P}$ at frames $t \in \tau$, individually designing each front-end for k phonetic classes might require considerable efforts in processing the obtained landmark streams, which might lead to fuzzy probability estimates, which will be discussed in the following paragraph.

Heterogeneous and asynchronous detection functions If each detection front-end uses different acoustic features, classifiers and post-processing, all k landmark sequences $x_k(\tau_k)$ correspond to heterogeneous scores that are not directly comparable. The use of heterogeneous acoustic observations for phonetic modeling is advocated in many studies, for example in the ASAT project (e.g., [SL09, BQH+07], see also section 3.2.2).

[1] Strictly speaking, the best hypothesis $q(t)$ at frame t corresponds to a state j of the state space $\mathcal{J}$ of the recognizer, with $q(t) \in \mathcal{J}$. This thesis uses the notation $q_{\mathcal{P}}(t)$ if the hypothesis is limited to the phone inventory $\mathcal{P}$ of the ASR system, so that $q_{\mathcal{P}}(t) \in \mathcal{P}$.

Yet, these approaches use k continuous detection functions, with the k scores $\{x_k(t)\}_k$ at each frame t serving as an input to a combination module, for example an ANN or CRF, that estimates an a posteriori probability for each phone at t. If each landmark detection front-end k is optimized individually, landmarks appear asynchronously, with $\tau_1 \neq \tau_2 \neq \ldots \neq \tau_k$. This makes retraining the scores at each frame t with an additional statistical classifier impossible. To embed asynchronous landmarks into speech processing, they are usually used as input sequences for statistical segmentation methods, for example in [JN09a, JEW08], where landmark sequences are converted into broad phonetic segments.

Obtaining phone posterior probabilities by using k detection functions as input to a combination module might result in fuzzy probability estimates, if the landmark detection strategy consists in detecting only very few phonetic classes, that do not cover the whole phone inventory. For example, if a landmark detection system detects only the best studied phonetic classes, like vowels, plosives and fricatives, and disregards phonetic classes which are known to have complex or subtle acoustic cues, like approximants and nasals, the acoustic observations and detection strategies used are most likely individually optimized for each phonetic class and thus not suitable to predict the presence of approximants or nasals.

Another reason for why a detection function $x_k(\tau_k)$ might be an unreliable input feature for additional statistical models, is the fact that modeling specific acoustic cues does often not account for all context-dependent cues of a phonetic class, for example by restricting vowels to local maxima of energy, which leads to unavoidable missed detections.

Overall one clearly sees that individually designed phonetic front-ends seem to have advantages concerning transferring phonetic concepts into accurate detection functions, yet one might want to sacrifice at least some degrees of freedom for the sake of simple post-processing and accurate probability estimates.

4.3. Binary landmarks

In the presented framework, each k-th detection function $x_k(\tau_k)$ is not directly integrated into the search for the best word hypothesis, but converted into a binary indicator $\Lambda_k(t)$ that indicates the time frames t when the phone of the correct hypothesis $q_{\mathcal{P}}(t)$ at t is supposed to be included in $\mathcal{S}_k$, with $\Lambda_k(t) = 1$. It is important to note that there are two possibilities to interpret $\Lambda_a(t) = 0$ for a knowledge source a at frame t:

- If $\Lambda_a(t) = 0$ but all other detection functions are equally 0, with $\sum_k \Lambda_k(t) = 0$, there is no knowledge about the correct path at t at all.
- If $\Lambda_a(t) = 0$ and there is another phonetic class b for which $\Lambda_b(t) = 1$, the landmarks suggest that the correct path $q_{\mathcal{P}}(t)$ is included in $\mathcal{S}_b$ and not in $\mathcal{S}_a$.

For each of the two different landmark detection architectures presented in the previous section there is a different strategy for converting $x_k(\tau_k)$ into binary landmarks $\Lambda_k(t)$.

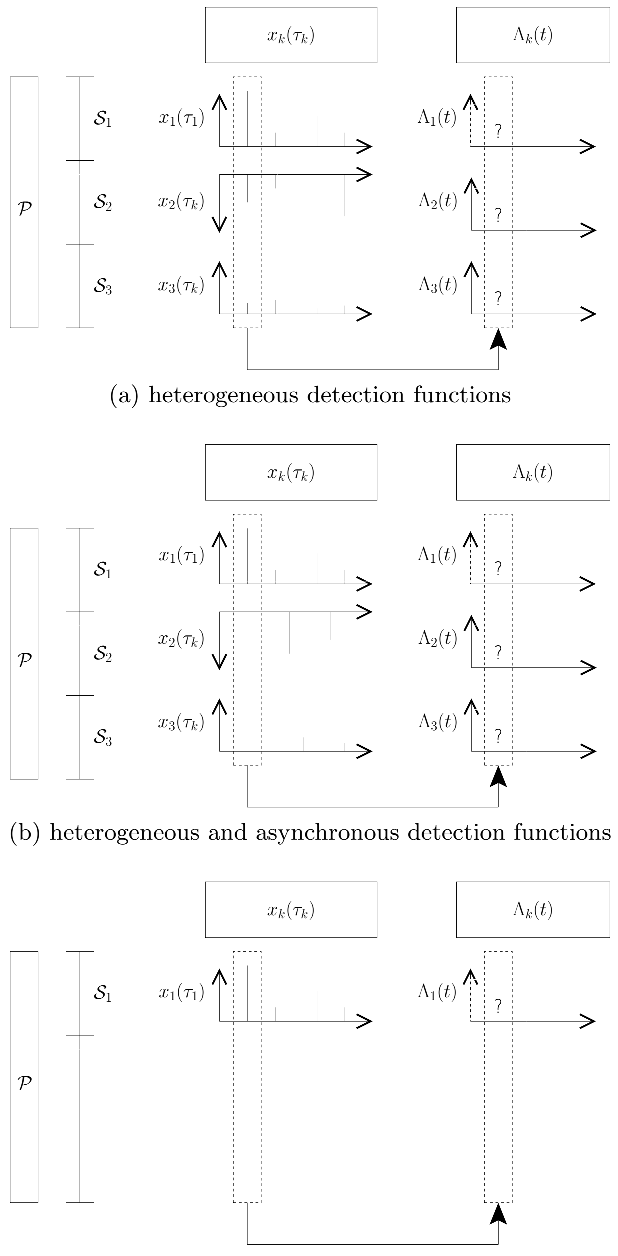
For a collection of individual detection frameworks, the strategy consists in extracting few, but very precise landmarks and one accepts to miss several phones of the correct word hypothesis. In this case, it might be sufficient to consider each detection function individually and pick landmarks by determining an individual threshold for each detection function $x_k(\tau_k)$, since landmark positions are supposed to be clearly indicated in the profile of $x_k(\tau_k)$. Additionally, there should be few to no overlap between potential binary landmarks of different phonetic classes and all values of $x_k(\tau_k)$ that are difficult to interpret or involve a certain degree of insecurity are discarded.

The second strategy, using a single front-end for all phonetic classes, aims at activating a binary landmark for each $t \in \tau$. Since it is desirable to get a complete landmark-based transcription of a speech utterance, each phone of the correct hypothesis is ideally associated with one time instance t. Since this will include acoustically ambiguous parts, each $x_k(t)$ at $t \in \tau$ should correspond to an accurate probability estimate about the presence of phonetic class k at frame t. Thus, detection errors, which are most likely to appear at frames where two or more phonetic classes obtain comparable prediction scores, are tolerated as long as the majority of predicted phonetic landmarks correctly match the best path.

4.4. Integration of landmarks into ASR

There are four basic techniques that are proposed in the literature to integrate external knowledge sources, i.e., detection functions, into HMM-based ASR, which are decision fusion, feature fusion, classifier combination and knowledge-based pruning (see for example [PNLM04, Met05, GM07]). Out of those approaches, decision and feature fusion are clearly not suitable to be employed for the integration of phonetic landmarks as they have been presented in the previous section. Decision fusion cannot be used, since the phonetic knowledge sources used in this thesis are not part of a standalone ASR system and decision fusion, for example using the ROVER approach [Fis97], combines the output hypotheses of two different systems. Feature fusion concatenates the acoustic observation vector $\mathbf{x}(t)$ and the k-dimensional continuous knowledge sources $\{x_k(t)\}_k$ at frame t to a new observation vector $\mathbf{x}'(t)$ to retrain the acoustic models with this extended observation vector. This integration technique is obviously not compatible with landmarks, which only are defined at certain time instances. The remaining two combination approaches that can account for landmark-based detection functions are classifier combination and phonetic pruning.

Classifier combination Classifier combination approaches aim at improving the classification performance of individual statistical classifiers by combining their prediction outputs into an improved common prediction score, thus benefiting from the complementary information that each classifier provides about the classification task. In speech recognition, classifier combination is used to combine the emission probabilities of the baseline with external predictions, for example k prediction scores $x_k(t)$. The dominant classifier combination method in ASR is the combination of logarithmic emission



(a) heterogeneous detection functions

(b) heterogeneous and asynchronous detection functions

(c) incomplete detection functions

Figure 4.3.: This figure illustrates the issues resulting from designing individual detection front-ends for each phonetic class when it comes to making predictions about the presence of phonetic classes, using the example of extracting binary landmarks from heterogeneous, asynchronous and incomplete knowledge sources $x_k(\tau_k)$. It can be seen that deciding whether to activate $\Lambda_k(t) = 1$ at frame t might be difficult, given the presented detection functions, since scores are not directly comparable (a), landmarks appear asynchronously (b) or only a fraction of the full phone alphabet is modeled (c).

probabilities and external knowledge via weighted linear combination.

One can clearly see that classifier combination favors the use of a shared front-end for k detection functions, as described in section 4.2, since combining classifier scores needs homogeneous and accurate probability estimates. Furthermore, classifier combination also favors the use of one common modeling technique for the acoustic models and the external knowledge sources, mostly corresponding to GMMs. To use classifier combination with landmark-based detection functions $x_k(\tau)$, rather than continuous detection functions $x_k(t)$, one has to use a model-change approach, i.e., classifier combination is only carried out at frames where landmarks are present, while the standard decoding is used otherwise.

Phonetic pruning As discussed in Chapter 2, there are several methods to reduce the search space during decoding, notably beam pruning, histogram pruning and language model pruning. [GM07] proposed an additional landmark-based pruning criterion, with active hypotheses at frame t getting pruned away if they are not compatible with binary landmarks that indicate the presence of phonetic classes. Since pruning the stack of active hypotheses is a severe modification of the search for the best word hypothesis, binary landmarks $\Lambda_k(t)$ have to be very precise (see following chapter 5)tge. Therefore, phonetic pruning is compatible with multiple independent phonetic detection front-ends, as described in Section 4.2, since only the most precise instances of $x_k(\tau_k)$ will be converted into binary landmarks $\Lambda_k(t) = 1$, which can be obtained by considering each detection function individually and noisy values of $x_k(\tau_k)$ will just be discarded, without the need of accurate probability estimates.

4.5. Summary and conclusions

This chapter introduced the general framework that converts the speech signal into a collection of binary landmarks that will be used to bias the search space in the following chapters. It could be seen that there are several simplifications concerning phonetic knowledge as it is defined in this thesis, for example by limiting the detected labels to natural speech classes or not considering the dependencies between phonetic landmarks. Yet, there is a considerable amount of freedom in adapting each landmark detector according to phonetic knowledge, if each detection function relies on its individual detection front-end.

To study whether it is possible to integrate such a collection of individual landmark detectors into the search of standard ASR, the next chapter will, as the first major contribution of this thesis, extend the experiments conducted in a previously proposed method for landmark-based pruning of the search space by testing the sensitivity of this approach towards phonetic confusions, which has not been done in prior art.

5. Search-space pruning using phonetic landmarks

As it has been concluded in the previous chapter, landmark-based pruning seems to allow the highest degree of freedom in designing phonetically motivated landmark detectors and is thus a good starting point to study the potential of landmark-driven ASR. The use of phonetic landmarks to prune the search space during the decoding was initially proposed in [GM07], together with a first set of experiments based on oracle landmarks, i.e., landmarks derived from reference alignments, to estimate the potential of landmark-based pruning. This chapter applies this pruning method to a realistic setting, by pruning the search space with landmarks obtained by statistical classification. To judge the potential of landmark-based pruning without depending on a specific landmark extraction algorithm, reference alignments are used to determine landmark positions, but binary landmarks are obtained according to the prediction scores of statistical classification.

5.1. Phonetic pruning using oracle landmarks

This section summarizes the phonetic pruning algorithm proposed in [GM07], including the motivation for using broad phonetic classes for pruning the search space and the main results of the experiments conducted.

Landmark-based pruning The Viterbi algorithm displayed in Equation 2.6 can be rewritten for frame t, using log-likelihoods instead of probabilities according to

$$Q(j,t) = \log p(\mathbf{y}_t \mid j) + \max_i \left\{Q(i,t-1) + \log a_{ij}\right\} \tag{5.1}$$

with j being the state in the decoding graph and t the frame index. a_{ij} represents the transition probability from state i to j, which also incorporates the language model at word transitions. $\log p(\mathbf{y}_t \mid j)$ corresponds to the likelihood of the acoustic observation $\mathbf{y}_t$ for state j. In [GM07] external knowledge sources incorporate knowledge about the best path in a binary mask $\lambda_j(t) \in \{0,1\}$. $\lambda_j(t) = 1$ indicates that state j is a potential active hypothesis at t and $\lambda_j(t) = 0$ indicates that state j is not compatible with the external knowledge at t and thus should not be part of the final word hypothesis. Clearly, $\lambda_j(t) = 1$ for all j and t in case there is no knowledge about the correct path at all. To account for $\lambda_j(t)$ during decoding, the static transition probability $\log a_{ij}$ in Equation 5.1 is changed into a function of t as follows:

$$\log a_{ij}(t) = \log a_{ij} - I_j(t) \tag{5.2}$$

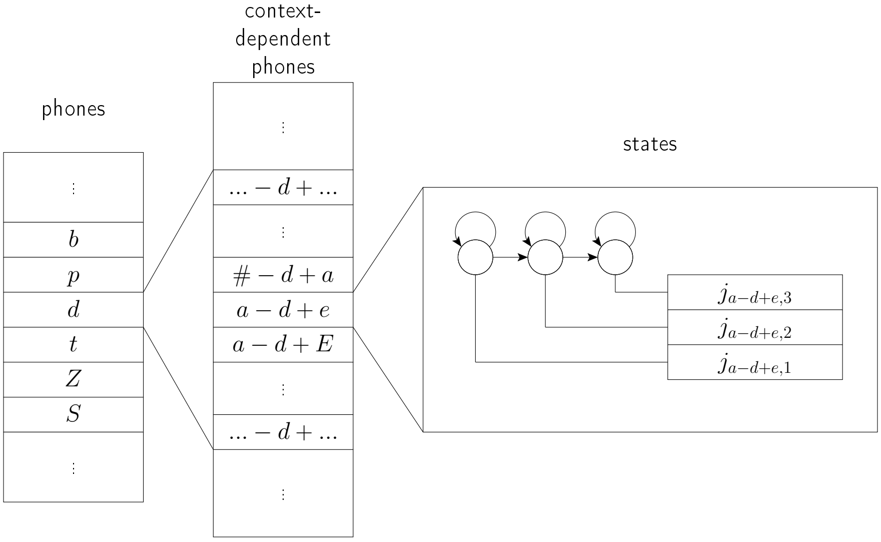


Figure 5.1.: Mapping of the phone d onto its context-dependent-phones and associated states of a three-state HMM.

with an indicator function $I_j(t)$ that simply turns to 0 if the state j is active according to the binary mask with $\lambda_j(t) = 1$ and to infinity if $\lambda_j(t) = 0$:

$$I_j(t) = \begin{cases} \infty & \text{if } \lambda_j(t) = 0 \\ 0 & \text{if } \lambda_j(t) = 1 \end{cases} \tag{5.3}$$

Thus, the decoding according to Equation 5.1 can be rewritten as

$$Q(j,t) = \log p(\mathbf{y}_t \mid j) + \max_i \{Q(i,t-1) + \log a_{ij}(t)\}. \tag{5.4}$$

This effectively cuts the transition to states j that are not compatible with the external knowledge at t according to $\lambda_j(t)$ and the external knowledge is used as additional pruning criterion for the stack of active hypotheses at frame t (see Figure 5.2).

While this method is a general algorithm for integrating knowledge into dynamic programming, it is straightforward to use this method in connection with the output $\Lambda_k(t)$ of the landmark detection framework proposed in the last chapter. First, the ensemble of phones $\mathcal{S}_k$, which is associated with each landmark detection function, has to be mapped onto its corresponding state inventory $\mathcal{J}_k$, by collecting all states associated with the context-dependent phones that are member of $\mathcal{S}_k$ (see Figure 5.1). Second, the binary indicator function $\Lambda_k(t) \in \{0,1\}$, with $\Lambda_k(t) = 1$ corresponding to landmarks indicating the presence of a phonetic class associated with phones $p \in \mathcal{S}_k$, has to be projected onto an individual state-based binary mask $\lambda_j^{(k)}(t)$ for detection function k by

$$\lambda_j^{(k)}(t) = \begin{cases} 0 & \text{if } \Lambda_k(t) = 1 \text{ and } j \notin \mathcal{J}_k \\ 1 & \text{else} \end{cases} \tag{5.5}$$

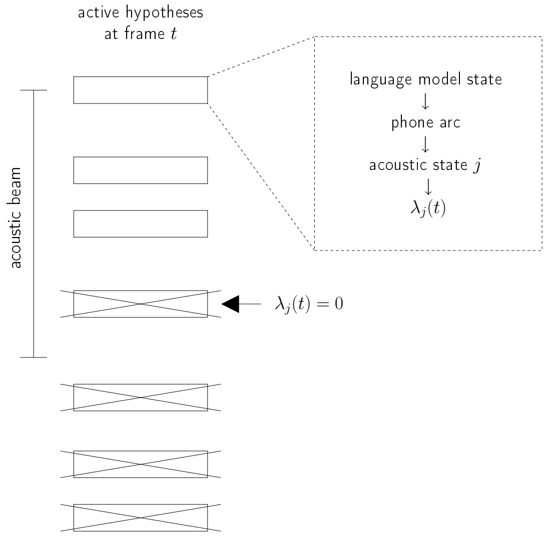


Figure 5.2.: Simplified example for using landmarks as additional pruning criterion by removing a hypothesis that is linked to a state j for which $\lambda_j(t) = 0$ from the stack of active hypotheses at time t. The uppermost hypothesis in the figure corresponds to the highest scoring active hypothesis and the distance between two hypotheses corresponds to the difference between the scores of these hypotheses.

The final binary mask $\lambda_j(t)$ is then obtained by a logical AND operation ($\bigwedge .$) over all k knowledge sources

$$\lambda_j(t) = \bigwedge_k \lambda_j^{(k)}(t) \,. \tag{5.6}$$

While $\lambda_j^{(k)}$ and consequently $\lambda_j(t)$ can be built dynamically along with the time-synchronous expansion of the search space, in the following it is assumed that $\Lambda_k(t)$ has been generated prior to the first pass of the speech recognizer.

Broad phonetic knowledge sources The phonetic classes k have to fulfill two requirements to effectively prune the search space as proposed in Equation 5.4. First, they have to be very precise since pruning away the correct path at frame t will have detrimental effects on the search for the best word hypothesis. Second, the knowledge sources must also be complementary enough to prune away paths that are not already pruned away due to the acoustic beam or might not be among the active hypotheses at t at all. [GM07] therefore proposes the use of binary landmark detectors indicating the presence of the broad phonetic classes vowels, semi-vowels, nasals, plosives and fricatives. BPCs have reported to be detectable with a high precision, since they correspond to the most distinctive articulatory gestures and their acoustic correlates are equally acoustically distinct. In [GM07], observing the stack of active hypothesis at each time frame t actually showed that the highest ranked hypotheses are often associated with

system	pass	baseline	BPCs	VPF	vow	plo	fri	nas	app
monophones	1	29.2	15.3	21.7	26.6	26.5	27.5	27.8	25.1
	2	22.3	13.9	17.6	21.2	20.7	21.0	21.5	20.1
triphones	1	27.3	19.6	23.9	27.0	26.3	26.0	26.4	24.9
	2	21.3	15.0	18.2	20.7	20.4	20.3	20.7	19.6

Table 5.1.: Reported word error rates after each pass for the two systems employed in [GM07] for landmark-driven decoding. «BPCs» corresponds to the use of all BPCs during decoding, «VPF» corresponds to the use of vowels, plosives and fricatives only and the remaining results correspond to the use of single BPCs.

different broad phonetic classes, which shows that correctly detected broad phonetic classes could improve the decoding by pruning incorrect hypotheses.

Experimental setup and oracle landmarks [GM07] used two baseline speech recognition systems, which are similar to the baseline system used in this thesis, with one system using context-independent models and another using word-internal context-dependent models. Both systems use two passes, generating a word graph in the first pass, which is rescored with more sophisticated acoustic models. The study proposes to use landmarks only in the first pass, since this allows to obtain improved word graphs for rescoring that already account for phonetic knowledge.

The speech data used for the recognition experiments were four hours of broadcast news taken from the development data of the ESTER broadcast news rich transcription evaluation campaign [GGG^{+}06]. All experiments conducted relied on landmarks that were derived from the force alignments of reference utterances, with each $\Lambda_k(t)$ being initialized by $\Lambda_k(t) = 0$ and only activated with $\Lambda_k(t) = 1$ at correct time instances. The width of the landmark, i.e., the number of frames per forced aligned phone that were activated with $\Lambda_k(t) = 1$ varied between one frame and several frames corresponding to maximal 70% of the overall segment length.

Results WERs obtained after both passes are summarized in Table 5.1. Both systems showed considerable improvement for the two different acoustic models after the first and second pass. With the use of oracle landmarks, the performance of the monophones gets very close to the performance of the triphone-based system, thus providing the potential of very fast decoding, since the search space of the monophone models is considerably smaller than the search space of the triphone models. The reduction in active hypotheses at each frame due to the pruned hypotheses was reported to reduce the decoding time by a factor of four.

While the improvement in WER could be expected since oracle knowledge was used, two observations indicated a possible improvement using realistic phonetic knowledge. First, even with few broad and well detectable classes (vowels, plosives and fricatives) improvement was considerable. Second, improvement was a linear function of the missed detection rate, since by randomly assigning only half of the speech units with a landmark (instead of all of them) the improvement in WER dropped 50%. That would still correspond to a considerable improvement in WER, given 50% of missed landmarks using only vowel, fricative and plosive landmarks.

Interestingly, experiments showed that there was no significant difference between pruning only one frame inside a force aligned subword unit or pruning away nearly all frames of this unit.

The natural question that arises from these experiments is whether phonetic pruning can cope with realistic landmarks that are subject to common acoustic confusions, which is examined in the following section.

5.2. Landmark-based pruning under realistic conditions

One drawback of the experiments in [GM07] was that oracle landmarks were selected without taking the difficulty of detecting the actual broad phonetic landmarks into account. Thus, landmarks correctly pruned hypotheses in very difficult parts of the speech signal, which are very unlikely to be correctly detected by statistical classifiers. To really judge the potential of this approach it is thus necessary to test it with automatically detected landmarks that are sensitive to acoustically ambiguous parts in the speech utterance.

Semi-oracle knowledge An intuitive extension of the described oracle experiments is to keep the same reference alignments, to obtain the landmark positions, but exchange the oracle landmarks with landmarks obtained according to statistical classification. Knowing the reference alignments is a considerable a priori knowledge which helps to obtain statistical predictions that are still much more precise than real world examples, yet they are subject to realistic confusions. So while improvement of the WER by using these «semi-oracle» landmarks would not necessary guarantee improvement for landmarks that make absolutely no use of an oracle, a degradation of the speech recognition performance caused by these landmarks would raise the legitimate question whether this approach is of practical use for LVCSR at all.

Landmark extraction Given a speech utterance, semi-oracle landmarks are extracted in three steps, which is also illustrated in Figure 5.3 with a simplified example.

1. Monophone acoustic models (see Appendix A.3) are used to force align the reference utterance in order to obtain a phone-level transcription of the utterance.

2. Each n-th segment, corresponding to an aligned phone, is reduced to a landmark t_n, by determining the temporal center of the segment, resulting in set of landmark positions $\tau = \{t_1, \ldots, t_n\}$. Each frame t_n is then associated with a score $x_k\left(t_n\right)$ for each context-independent phone k, by force aligning the corresponding HMMs with the segment. This might be seen as a special case of knowledge sources

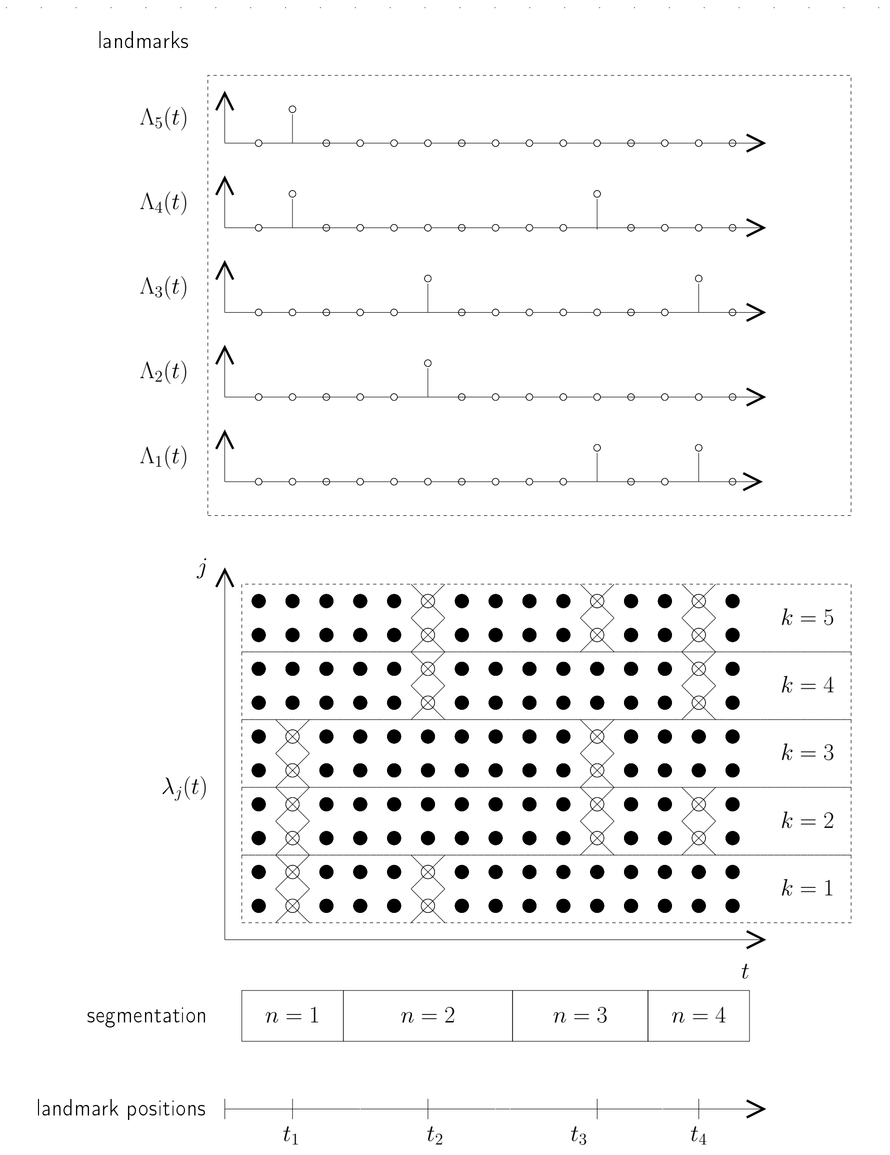


Figure 5.3.: Simplified example for the use of monophone models to extract binary landmarks to obtain a binary mask $\lambda_j(t)$. The binary mask is illustrated as filled dots for $\lambda_j(t) = 1$ and empty dots for $\lambda_j(t) = 0$. The pruned transitions to states for which $\lambda_j(t) = 0$ are displayed as pruned states for simplicity, displayed as crossed out dots. The example includes $k = 5$ phones each phone with two states and $n = 4$ force aligned segments. The $m = 2$ highest scoring phones are kept at each segment which for $n = 1$ corresponds to $k = 5$ and $k = 4$, for $n = 2$ to $k = 2$ and $k = 3$, for $n = 3$ to $k = 1$ and $k = 3$ and for $n = 4$ to $k = 1$ and $k = 3$.

m	baseline	18	16	12	8
WER	40.5	41	41.8	44.3	51.6

Table 5.2.: WER after two passes using landmark-based pruning with different sizes of the merging parameter m at a constant size of the acoustic beam. The lower m, the higher the number of erroneous landmarks that are used for pruning. The highest possible value of m is 39, which would correspond to no pruning at all.

according to Chapter 4, with phonetic class k and the associated set of phones $\mathcal{S}_k$ corresponding to a single phone.

3. At each landmark t_n, the m highest scoring phones are kept to activate an individual set of m phones at each landmark. Thus, the detection functions $x_k(\tau_k)$ are converted into binary landmarks $\Lambda_k(t)$, with $\Lambda_k(t) = 0$ for all $t \notin \tau$ and $\Lambda_k(t) = 1$ if phone k is among the m highest scores at t_n.

Merging the m best monophones into one speech class at each landmark allows to influence the number of states that are pruned away at each landmark. This can vary from many states for low values of m, which bears the potential to prune many incorrect hypotheses during the decoding, to few states that will prune less active hypotheses, while equally introducing less errors into the landmark-driven decoding. Since the experiment is designed to focus on acoustic confusions in general, the experiments are based on monophone models instead of specifically trained broad phonetic models.

Experiments The experimental setup is similar to the one in the original study in the sense that a similar two-pass triphone system is employed. The speech recognition results presented in the following have been conducted on a small subset of the ESTER 2 development set (see Appendix A.2).

Additionally to changing the merging parameter m, landmark-driven decoding was carried out with different sizes of the acoustic beam, to study the relation between landmark errors, decoding speed and WER. Speed is measured as the average number of hypotheses per frame, resulting from the remaining hypotheses at each frame after acoustic and language model pruning, as well as landmark-based pruning have been conducted. Comparing landmark-driven decoding and the baseline at different beam sizes should reveal whether using landmarks can at one point lead to faster decoding at equal WER compared to the baseline.

Table 5.2 displays the WER of the baseline system compared to the landmark-driven decoding using 4 different values for m. It can be seen that landmark-driven decoding performs considerably worse than the baseline with decreasing m. Even for very broad landmarks with $m = 18$, the WER is still 0.5% higher than the baseline. Fig 5.4 displays the relation of one landmark-driven system that caused only minor degradation of the WER ($m = 16$) and the baseline using varying sizes of the acoustic beam. Again, landmark-driven decoding cannot compete with the baseline ASR system, since the

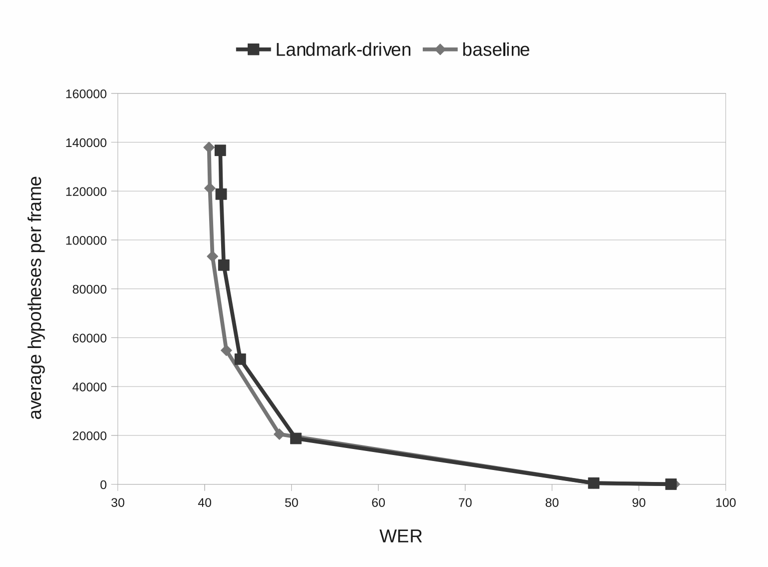


Figure 5.4.: WER of the baseline and landmark-driven decoding (using $m = 16$) as a function of average hypotheses per frame.

same decoding speed of the landmark-driven system can always be achieved by the baseline system at a better WER.

5.3. Discussion

While the landmark extraction algorithm used in this experiments had knowledge about the correct position of the landmarks according to reference alingements, the landmark-driven decoding did not outperform the baseline ASR system, regardless whether landmarks pruned many or only few states of the search space. The poor performance of the landmark-based pruning can be due to two factors. First, external knowledge sources will inevitably contain false landmarks due to acoustic confusions, which will prune away the correct path in the output lattice of the first pass, inevitably leading to word errors in the second pass. Second, increasing the precision of the landmarks comes at the expense of more missed landmarks. If the selection of landmarks is limited to the most confident landmarks, as it is the case for $m = 18$, the landmarks are unlikely to provide additional knowledge to the acoustic models, since the states that will get pruned away do not correspond to active hypotheses or will be correctly rescored in the second pass, no matter which word graph is provided by the first pass.

While these results seem discouraging for the use of phonetic landmarks in connection with phonetic pruning, there are several possible additional experiments worth studying in future work. One might consider additional experiments on a less difficult task, including clean environments, less speakers and less variation concerning channels and pronunciation, to test whether phonetic landmarks can provide benefits concerning speed or word accuracy when there is less acoustic variability. Since the provided phonetic landmarks are still very similar to the acoustic models, they are likely to be sensitive to similar acoustic confusions and capture the same acoustic properties. Thus, the pruning scheme might benefit from more complementary knowledge, like it can be provided for example by the visual modality. One possible approach might be the detection of visual gestures as single time instances, which are likely to be very complementary to the acoustic signal, but very reliable to detect.

The experiments clearly indicated that landmark-based pruning in general might either need very precise landmarks or very complementary knowledge to be effective. Since both assumptions are unrealistic to achieve in the short term, the pragmatic conclusion from the obtained results is to relax the integration method to allow less precise landmark modeling, which will be one of the contributions of the next chapter.

6. Guiding search in ASR by broad phonetic landmarks

This chapter provides two contributions on the road towards ASR driven by phonetic landmarks. First, instead of using landmarks as an additional pruning criterion to reduce the search space, this chapter introduces landmarks as additional score into the Viterbi decoding by multiplying binary landmarks with a fixed weight to combine them with the emission probabilities of the acoustic models. This strongly reduces the pruning of correct hypotheses due to erroneous landmarks, since landmarks will change the score of the active hypotheses rather than remove them completely. Second, it presents the first landmark detection framework that has been developed during this thesis, which is based on a segmentation and classification approach, i.e., the speech signal is segmented into short segments, with segment boundaries indicating major changes in the articulatory movements, before acoustic information is extracted from each segment to predict the associated broad phonetic class. The acoustic information corresponds to concatenated spectral observations of subsequent segments to provide information about dynamic changes in the articulatory gestures to obtain robust landmarks that are more reliable than conventional frame-based acoustic predictions.

6.1. Modified Viterbi decoding

As stated in the previous chapter, it is desirable to replace landmark-based pruning by a different implementation, which allows to insert erroneous landmarks into the decoding that do not inevitably lead to errors during the search for the best word hypothesis. An intuitive change of Equation 5.4 consists in replacing hard pruning with a combination approach, that modifies the score of active hypotheses rather than pruning paths completely.

This new combination method also relies on external knowledge, represented as a binary mask $\lambda_j(t)$, but adds a fixed score to states j for which $\lambda_j(t) = 1$ instead of pruning the states that are incompatible with this knowledge. It is important to note that while in the previous chapter the binary mask $\lambda_j(t)$ was $\lambda_j(t) = 1$ by default, i.e., if no external knowledge is present, the new combination method requires $\lambda_j(t) = 0$ if decoding is run without any knowledge. Thus, the step from binary landmarks $\Lambda_k(t)$ to binary mask $\lambda_j^{(k)}(t)$ corresponds to the inverse procedure of Equation 5.5 in Chapter 5 (see also Figure 6.1 for a simplified example):

$$\lambda_j^{(k)}(t) = \begin{cases} 1 & \text{if } \Lambda_k(t) = 1 \text{ and } j \in \mathcal{J}_k \\ 0 & \text{else} \end{cases} \tag{6.1}$$

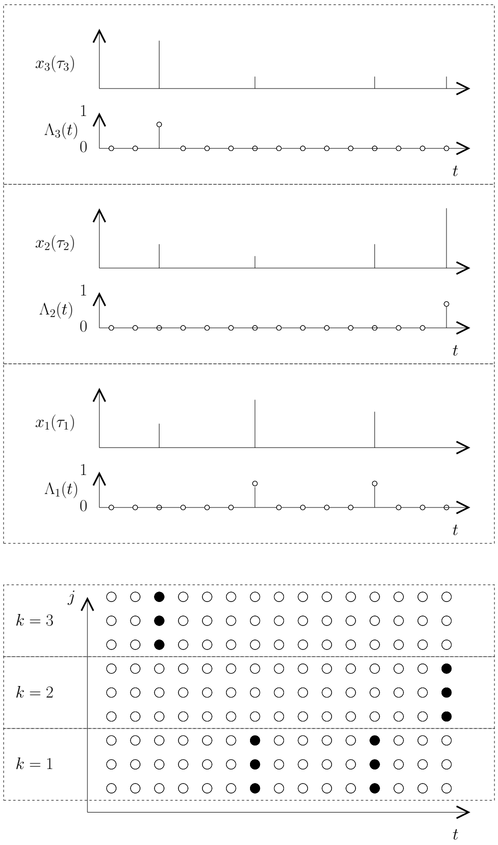


Figure 6.1.: Simplified example of $k = 3$ detection functions $x_k(\tau)$, with $\tau = \tau_1 = \tau_2 = \tau_3$, and extracted landmarks $\Lambda_k(t)$, which are further converted into the binary mask $\lambda_j(t)$, assuming each phonetic class k is associated with 3 states. Active states $\lambda_j(t) = 1$ are displayed as filled dots and the highest ranked class at each landmark τ is activated with $\Lambda_k(t) = 1$.

and the final binary mask is obtained by a logical OR operation ($\bigvee_{\cdot}$) over all k knowledge sources

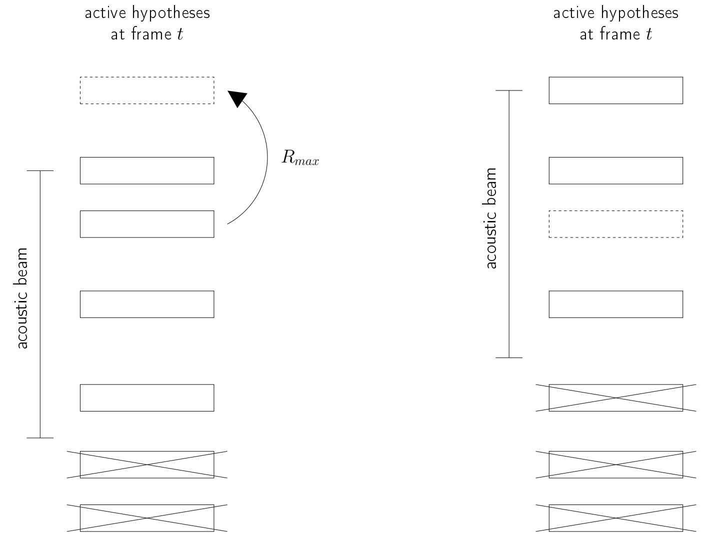
$$\lambda_j(t) = \bigvee_k \lambda_j^{(k)}(t). \tag{6.2}$$

The modified Viterbi decoding uses the binary mask $\lambda_j(t)$ as a third source of statistical knowledge, besides language and acoustic model, to keep track of the best hypotheses at t as follows:

$$Q(j,t) = \log p(\mathbf{y}_t \mid j) + \max_i \{Q(i,t-1) + \log a_{ij}\} + \lambda_j(t) R_{max} \tag{6.3}$$

It can be seen that Equation 6.3 corresponds to standard Viterbi decoding according to Equation 5.1 in case $\lambda_j(t) = 0$. With $\lambda_j(t)$ indicating whether state j is compatible with the external knowledge, R_{max} is the weight limiting the influence of each binary landmark $\lambda_j(t) = 1$ on the overall score.

This modified decoding aims at pruning incorrect hypotheses during the first pass, but since Equation 6.3 enhances the scores of states associated with the presumably correct solution instead of pruning presumably incorrect hypotheses, this can only lead to improved word graphs if the modification causes incorrect hypotheses to fall under the acoustic beam during the search process. This method is less beneficial if the knowledge sources do not contain errors and R_{max} is not high enough to put competing hypotheses under the beam. Yet, it is beneficial if some landmarks are erroneous and do not cause the pruning of the correct path, since the hypothesis will survive the acoustic pruning despite the weight being allocated to the wrong states in the search space. Thus, the enhancement factor R_{max} is the parameter mediating between the desire to prune away incorrect hypotheses and the need to avoid errors. Since only one parameter has to be estimated, R_{max} is directly obtained by optimizing the WER of the development set using a one-dimensional grid-search.



(a) before

(b) after

Figure 6.2.: The effect of reordering the stack of active hypotheses by modifying the score of the Viterbi decoding using $\lambda_j\,(t)R_{max}$.

Landmark detection front-end Since binary landmarks used inside landmark-based pruning directly modified the stack of active hypotheses, each activated binary landmark $\Lambda_k\,(t) = 1$ directly propagated into the final word hypothesis of the decoder, respectively the n-best list obtained after the first pass. In landmark-driven decoding according to Equation 6.3, each landmark only modifies the scores of the stack of active hypotheses. Therefore, the number of landmarks, i.e., the time instances where $\Lambda_k\,(t) = 1$, should be increased compared to landmark-based pruning, since landmarks have to occur frequently enough to result in improved word graphs. Since increasing the number of landmarks will include parts of the speech signal with high acoustic ambiguity, $x_k\,(t)$ needs to provide accurate probability estimates to activate binary landmarks.

Therefore, this chapter uses a single-front end based on a multi-class classifier to produce k homogeneous sporadic detection functions $x_k(\tau)$. The phonetic classes cover the whole phone inventory with $\bigcup_k \mathcal{S}_k = \mathcal{P}$ and landmarks appear synchronous with $\tau_k = \tau$, as it has already been discussed in Chapter 4. The detection functions $x_k(\tau)$ indicate the $k = 6$ broad phonetic classes vowels, semi-vowels, nasals, plosives, fricatives and a general class for all non-speech symbols of the phone inventory of the ASR system. The algorithm exploits the fact that these classes correspond to the most distinct articulatory gestures by embedding classification and detection in an articulatory motivated segmentation framework to obtain robust acoustic observations for classification.

6.2. Landmark detection

Following the general outline of the single front-end for landmark detection according to Chapter 4, the proposed algorithm detects a sequence of landmarks $\tau = t_1, \ldots, t_n$ and a statistical classifier predicts the probability of each k-th phonetic class for each frame t_n to obtain k detection functions $x_k(\tau)$, before k binary landmark indication functions $\Lambda_k(t)$ are extracted. The following paragraph will give an overview of the algorithm, while the details of each individual step are further explained in the following sections.

1. The speech signal $\mathbf{Y}$, consisting of parametrized frames $\mathbf{Y} = \mathbf{y}_1, \ldots \mathbf{y}_t$, is segmented into a sequence of n non-overlapping segments, with each segment corresponding to a sequence of frames $\left\{\mathbf{Y}_{a_n}^{b_n}\right\}_n$, with the n-th segment $\mathbf{Y}_{a_n}^{b_n}$ including the sequence of frames from frame index a_n to b_n. Each segment is supposed to correspond to a short part of speech where the spectrogram is nearly constant and the positions of the articulators are supposed to exhibit very low dynamics. The segmentation is unsupervised and can thus be equally applied to speech utterances for training and prediction. (see Section 6.2.1)

2. The spectral content of each segment is approximated by a fixed-dimensional observation vector $\mathbf{x}'_n$ and this observation vector is additionally augmented by adding its immediate neighbors as contextual information to obtain the final observation $\mathbf{x}_n = \left[\mathbf{x}'_{n-1}, \mathbf{x}'_n, \mathbf{x}'_{n+1}\right]$ for the n-th segment. This observation vector is either used to train a multi-class statistical classifier in the training phase or to predict segments of unseen speech signals to obtain k classification scores $X_k(n)$ for each segment. In the presented framework, the classification method corresponds to bagged decision trees. (see Section 6.2.2)

3. After classifying each segment, one frame inside the segment is determined as the landmark t_n to obtain a sequence of landmarks $\tau = t_1, \ldots, t_n$. Each classification score $X_k(n)$ is thus converted to $x_k(t_n)$ in order to obtain the final k detection functions $x_k(\tau)$. (see Section 6.2.3)

4. The k knowledge sources $x_k(\tau)$ are used to extract k landmark indication functions $\Lambda_k(t)$ by activating one or several sources k at each $t \in \tau$, according to a merging threshold δ.

While each phonetic class is not modeled by individual acoustic cues, the algorithm aims at using phonetically motivated models for segmentation and phonetically motivated acoustic features $\mathbf{y}_t$. Additionally, decision trees are used to obtain a rule-based classification similar to classification in acoustic-phonetic approaches to ASR.

The next section explains the motivation and details of the segmentation algorithm used, which can be considered to be the core of the proposed landmark detection framework, since the main difference between standard frame-based acoustic observations and the proposed segment-based acoustic observations is the fact that frame-based observations rely on fixed-size temporal windows, in contrast to the variable-size temporal windows obtained from segmentation.

6.2.1. Segmentation

Speech segmentation structures the speech signal into a sequence of n segments, each segment corresponding to a tuple (a_n, b_n) that indicates the index of the first and last frame of segment n. Unlike most segmentation algorithms, the desired segmentation does not have to correspond to a citation-phonemic segmentation of the speech signal, since the goal is not to detect phones, but to obtain a structured speech representation that adapts to the dynamics of the speech utterance to track the movements of the articulators. Since the phonetic targets of the k detection functions correspond to the major articulatory gestures, the sequence of segments should allow to capture the spectral evolution resulting from articulatory movements. There are several speech segmentation algorithms motivated by articulatory phonetics that associate spectral changes in the signal with changes in the articulatory filter (e.g., [Ata83] [BCDM88, VDKM89]). The segmentation method used in this chapter, which is described in [AO88], is based on a refined version of spectral change detection using autoregressive (AR) models which detects changes in the spectrum of a non-stationary signal by comparing the output of two AR models (see for example [BB83]). This segmentation model fits well for this task, since modeling speech as the output of a time-varying acoustic filter is very close to the source-filter model of human speech production.

Segmentation for landmark detection By modeling the acoustic waveform as an excitation signal modified by a time-varying autoregressive filter, changes in the filter parameters correlate with changes in the articulatory gestures. Given the general autoregressive model[1]

$$y_t = \sum_i a_i y_{t-i} + e_t$$

with the AR parameters corresponding to $A = a_1,, \ldots, a_i$ and the samples corresponding to $Y = y_{t-1}, \ldots, y_{t-i}$. In [AO88], the AR parameters A_0 of a growing long-term window and the AR parameters A_1 of a short-term window attached to the long-term window

[1] It should be noted that the notation in this paragraph is not synchronized with the remainder of this chapter.

articulatory feature	knowledge-based parameters
silent	E[0,F3-1000], E[F3,$f_s/2$], ratio of spectral peak in [0,400Hz] to the spectral peak in [400,$f_s/2$], Energy onset, Energy offset]
sonorant	E[0,F3-1000], E[F3,$f_s/2$], Ratio of E[0,F3-1000] to E[F3-1000,$f_s/2$], E[100,400]
syllabic	E[640,2800], E[2000,3000], Energy peak in [0,900Hz], Location in Hz of peak in [0,900Hz]
continuant	Energy onset, Energy offset, E[0,F3-1000], E[F3-1000,$f_s/2$]

Table 6.1.: Example for knowledge-based APs, directly taken from [EWPJD07]. The description of the parameters relies on the following annotations: f_s (sampling rate), F3 (third formant average), [a,b] (frequency band [Hz,Hz]), E[a,b] (energy in the frequency band [Hz,Hz]).

are calculated to determine changes in the signal by calculating the cumulative sum of the conditional cross-entropy between the two models, which is further coupled to the Page-Hinkley test to obtain clear indications of the spectral changes in the signal. To improve robustness against omissions, change detection can be additionally carried out in the backward direction. Since this segmentation method is done on a per-sample basis, obtained segment boundaries are mapped onto the nearest frames.

While this algorithm has been originally proposed to obtain a citation-phonemic segmentation, the algorithm can be tuned to oversegment speech units so that the resulting segments divide speech into spectrally stable or slowly transitional units of variable lengths. The main parameters that have to be tuned are the parameters of the Page-Hinkley test, the model order i and the length of the short-term observation window. The parameters used for the experiments in Section 6.3 are hand-tuned, i.e., the parameters were chosen so that the obtained segmentation of selected utterances fitted the acoustic changes visible in the spectrogram.

While the obtained speech segments are in the first place used for landmark detection, they are also employed during the feature selection process that determines which acoustic observations are included in the speech parametrization $\mathbf{y}_t$ for each frame, which is described in the following section.

6.2.2. Frame and segment-based acoustic observations

Instead of a general frame-based representation of the spectrogram, like MFCCs, landmark classification and detection are based on acoustic features that are adapted to the k broad phonetic target classes.

Knowledge-based acoustic features BPCs correspond to the most discriminative articulatory gestures and the corresponding acoustic cues are supposed to be detected by the human auditory system using several coarse measurement operations during early stages of speech perception, which are then combined to estimate higher level linguistic units ([Ste02]). To mimic this process, the presented algorithm uses m spectral measurements, sometimes referred to as acoustic parameters (APs), at each time frame t as input features with $\mathbf{y}_t = y_{1,t} \ldots, y_{m,t}$. In related work (for example [EWPJD07]), each AP corresponds to a measurement score obtained by processing spectral bands to determine energy, energy rations, spectral peaks and onsets or offsets, as they are for example displayed in Table 6.1. The boundaries of frequency bands are either derived by manually observing the spectrogram or rely on the estimation of the formant frequency. Yet, manually determined frequency bands might be very speaker dependent and formants result in noisy measurements for highly variable speech environments. To use APs in more difficult recognition tasks, it is necessary to rely on simpler, yet more robust acoustic measurements as acoustic features $y_{m,t}$. Since the number of possible acoustic features that can be used as APs is very high, this landmark detection framework relies on feature selection to select an optimal set of features out of a pool of potential acoustic features.

Correlation-based feature selection of acoustic features The pool of potential acoustic features used in this thesis consists of a huge amount of features that are frequently involved in knowledge-based parameters, like high-frequency content measures, different spectral energies, on and offset detectors, formant energy and frequency and other pitch related parameters, which are listed in detail in Appendix A.3.2. To determine the optimal set of relevant acoustic features, feature selection has to choose the ensemble of features that provides an optimal combined prediction power. In this work, correlation-based feature selection (CFS) is used, based on the work in [Hal99], which performs well in several selection tasks [HH03]. CFS selects features based on evaluating multi-dimensional feature subsets according to a selection criterion that takes into account the predictive power of the feature subset, as the average class-feature correlation of the subset, as well as the average inter-feature correlation of the subset. This criterion results in high scores if the subset under consideration correlates well with the associated classes, while the features are not correlated among each other.

Given n speech segments obtained from the segmentation method described in Section 6.2.1 as a tuple $\{(a_n, b_n)\}_n$, corresponding to first frame a_n and last frame b_n of the segment, feature selection is performed in four steps:

1. The temporal center $t_n^{(c)}$ of each segment n is selected, formally calculated as the nearest integer to $a_n + \frac{b_n - a_n}{2}$.

2. For each frame $t_n^{(c)}$, the initial pool of M features is selected, resulting in one feature vector $\mathbf{y}_t^{(M)}$ for each segment.

3. As a purely segment based feature, the length of each segment $l_n = b_n - a_n + 1$ is also added to $\mathbf{y}_t^{(M)}$.

selected acoustic feature	selected bands or coefficients
zero crossing rate	
energy	
octave band signal intensity (OBSI)	7,3,4,5,6,8,2
log-ratio of consecutive OBSI	2,1,4,4,6,7
spectral crest factor per band	16
MFCC	12,13,25,26,38,0,4,19,18,20,21,22,23,24,6,15
temporal shape statistics	0,2
normalized bark bands	0,4,5,23,2,3,18,20,21,22,6,8
spectral variation	
spectral shape statistics	0,1,2,3
spectral deviation	
spectral roll-off	
spectral flatness per band	13
formant frequency	0,[1,2,3]
pitch	
formant amplitude	0,2
length of the segment	

Table 6.2.: Final set of selected features, including information about the indices of multi-band or multi-value acoustic features. While only the first formant frequency was initially selected, all four formant frequencies have been added to the set of final features.

4. The CFS algorithm uses the collection of n feature vectors $\left\{\mathbf{y}_t^{(M)}\right\}_n$ as well as the associated broad phonetic labels $\{c_k\}_n$ to obtain the final set of m acoustic features.

Selected features Out of all features, many of them being multidimensional features, correlation-based feature selection resulted in $m = 62$ features, displayed in Table 6.2. Zero crossing rate, a measure for high frequency content is the highest ranked feature, possibly due to its high correlation with fricatives. Following are acoustic features that correspond to rather broad spectral energy or energy ratios and several selected MFCC coefficients, mostly high frequency coefficients of first and second order. This shows that broad spectral measurements might be more discriminative, compared to the relatively fine spectral distinctions of MFCC coefficients, when it comes to distinguishing the basic articulatory movements. An additional interesting observation is the low ranking of formants, where only the first formant has been selected, and pitch related acoustic features, which are probably too speaker dependent. Yet, since a formant frequency is usually only meaningful when considered in context with other formant frequencies, all formants are added to the final set of acoustic features. The length of the segments is ranked at the very last, which shows that adding temporal information to acoustic measurements is a highly variable information that is not discriminative for different phonetic classes.

Normalization Since acoustic features are to a certain degree speaker and channel dependent, features have to be normalized onto a common range to allow reliable predictions. Given information about the speaker identity from the diarization step of the speech recognition framework, the following experiments will make use of two different normalization techniques. Besides the standard mean and variance normalization, normalization is also performed by quantile normalization or Equal Interval Width Binning [DKS95]. This normalization technique uses the q quantiles of the distribution obtained by ordering the t features $y_{m,t}$ from lowest to highest value before dividing the ordered features into q equally-sized subsets, corresponding to the q quantiles. The value of each feature $y_{m,t}$ is then replaced with the corresponding quantile q to obtain a normalized value $\hat{y}_{m,t}$.

The following paragraph concludes the discussion about acoustic observations by presenting the step from frame to segment-based acoustic observations.

Acoustic observations Algorithm 6.1 gives a detailed overview of the extraction of segment-based acoustic observations. The first step consists in extracting an observation vector $\mathbf{x}'_n$ from each segment $\mathbf{Y}_{a_n}^{b_n}$, before following and preceding observation vectors are concatenated to the final vector $\mathbf{x}_n = \left[\mathbf{x}'_{n-1}, \mathbf{x}'_n, \mathbf{x}'_{n+1}\right]$ for each segment to provide information about the temporal context of the acoustic observations. $\mathbf{x}'_n$ approximates the acoustic content of each segment by its central three observation frames, since the center of a segment is supposed to contain spectral information that is the least influenced by articulatory transitions and thus providing the most invariant short-term acoustic information about the acoustic correlate of the articulatory movement. The length of each segment, corresponding to an integer value representing the number of frames in each segment, is then added as an additional feature to the observation vector. $\mathbf{x}_n$ is used to a train a segment-based statistical classifier in the training case or to predict the label of unlabeled segments otherwise.

6.2.3. Classification and landmark extraction

The two steps in the landmark detection framework that remain to be discussed are the statistical classifier that estimates the probability of the k phonetic classes and the extraction of binary landmarks $\Lambda_k(t)$.

Decision trees to learn phonetic rules The proposed landmark detection framework uses decision trees as statistical classification method $F(\mathbf{x}_n)$ to predict $X_k(n)$ given $\mathbf{x}_n$, with $X_k(n) = F_k(\mathbf{x}_n)$. A node in a decision tree corresponds to a question concerning the value of one of the m input features, providing two branches leading to the next layer in the tree for each possible answer until a final node is reached that provides probability estimates for all k classes. Choosing decision trees over a variety of other classifiers is supposed to be a compromise between acoustic-phonetic approaches to ASR,

Algorithm 6.1 Predicting the phonetic class for speech segments.

```
Input: $\left\{\mathbf{Y}_{a_n}^{b_n}\right\}_n = \{\mathbf{y}_{a_n}, \ldots, \mathbf{y}_{b_n}\}_n$

for each $n$-th segment $\mathbf{Y}_{a_n}^{b_n}$:

    1. Determine the center of the segment $t_n^{(c)} = a_n - 1 + \frac{b_n - a_n}{2}$

    2. Extract $3 \times m$ acoustic features, with $m$ corresponding to the dimensionality of $\mathbf{y}_t$:

       $\mathbf{x}'_n = \left[\mathbf{y}_{t_n^{(c)}-1}, \mathbf{y}_{t_n^{(c)}}, \mathbf{y}_{t_n^{(c)}+1}\right]$

    3. Add the length of the segment $l_n$ to the observation vector $\mathbf{x}'_n$

    4. Concatenate the observation vector with its neighbors:

       $\mathbf{x}_n = \left[\mathbf{x}'_{n-1}, \mathbf{x}'_n, \mathbf{x}'_{n+1}\right]$

    5. The prediction score $X_k(n)$ for segment $n$ and phonetic class $k$ is obtained by using $\mathbf{x}_n$ as the input of the trained classifier $F_k(\mathbf{x}_n)$:

       $X_k(n) = F_k(\mathbf{x}_n)$

Output: $X_k(n)$
```

that identify acoustic cues by a set of phonetic rules derived by experts, and the need for statistical classifiers that can automatically learn the relation between $\mathbf{x}_n$ and a phonetic class k to account for the variability of speech. Thus, the resulting tree can be viewed as a set of measurements and corresponding thresholds, as they are common in acoustic-phonetic speech processing, yet these measurements are automatically derived and their complexity reflects the complexity of the speech process. An additional advantage is the fact that decision trees perform an intrinsic feature selection, since following the top node to the final node of the tree evaluates phonetic classes by an individual chain of measurements. Ideally, one expects general measurements at the top of the tree, that account for the general differences between articulatory gestures and more specific measurements towards the bottom, that account for all kinds of speech variability. Probability estimates $X_k(n)$ for each class k are based on frequency counts of training instances ending up in the final nodes. The final advantage of decision trees is that they accommodate for nominal information, which allows to add the channel bandwidth, obtained from the diarization step of the speech recognition system, as an additional feature in the experiments conducted in the next section.

The trees are trained by continuously splitting the training instances at each node according to an optimality criterion judging the quality of the split. This algorithm uses information gain as the splitting criterion to generate a maximum size tree and a held-out set of the training data is used to prune away branches that overfit the training data. Bootstrap-aggregating (bagging) is used as an additional tool to avoid overfitting. Bagging is performed by using i random subsets of the training data, containing a fixed amount of bootstrap samples of the data to create i decision trees. The prediction score is then averaged over all individual predictions.

Segment and frame-based classifier Decision trees are used to train two different classifiers. The first classifier $F_k(\mathbf{x}_n)$ provides a prediction score $X_k(n)$ for each segment-based feature vector $\mathbf{x}_n$ and the second classifier $f_k(\mathbf{y}'_t)$ provides a prediction score for each frame t, taking the frame-based observations $\mathbf{y}'_t = [\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{y}_{t+1}]$ as input. Training a frame-based BPC classifier has two main purposes. First, this classifier is used to obtain an estimate for the landmark position t_n for each segment n. Second, it is also used to benchmark the landmark-driven ASR system.

With the segment-based classifier being trained using all segment-based observations of the n segments $\mathbf{x}_n$ of every training utterance, the frame-based classifier is trained using the same number of training instances by using the central frame $t_n^{(c)}$ and corresponding observation vector $\mathbf{y}'_{t_n^{(c)}} = \left[\mathbf{y}_{t_n^{(c)}-1}, \mathbf{y}_{t_n^{(c)}}, \mathbf{y}_{t_n^{(c)}+1}\right]$ of each training segment. The output of the frame-based classifier results in standard frame-based continuous detection functions, with $x_k^{(\mathbf{y})}(t) = f_k(\mathbf{y}'_t)$. By using decision trees, both classifiers $F_k(\mathbf{x}_n)$ and $f_k(\mathbf{y}'_t)$ provide prediction scores $X_k(n) \in [0,1]$ and $x_k^{(\mathbf{y})}(t) \in [0,1]$.

Detecting the landmark position To determine the landmark position t_n of each segment n, which stretches from t_{a_n} to t_{b_n}, the frame based classifier $f_k(\mathbf{y'}_t)$ is used to obtain the frame where the class k with the highest segment score has its maximum frame score, which can formally be computed as follows:

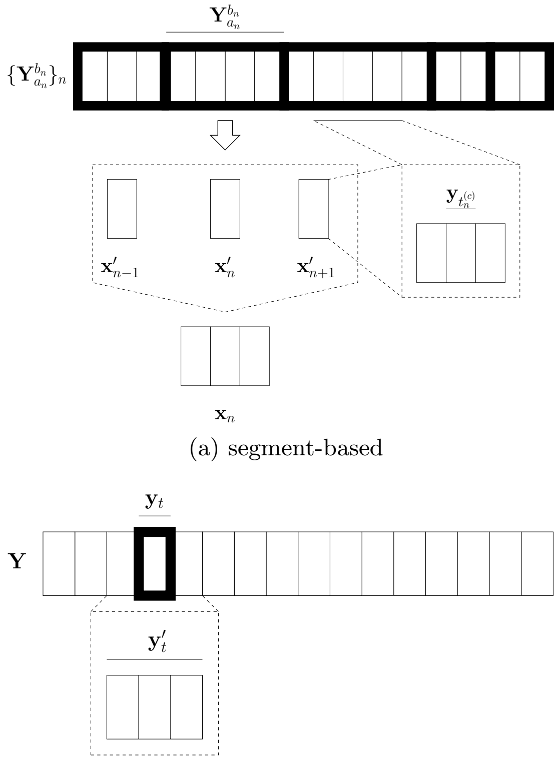
$$\hat{k}(n) = \arg\max_{k} X_k(n) \tag{6.4}$$

$$t_n = \arg\max_{t\in[t_{a_n}, t_{b_n}]} f_{\hat{k}(n)}(\mathbf{y}'_t) \tag{6.5}$$

with $\hat{k}(n)$ corresponding to the class k at segment n with the highest score $X_k(n)$. After determining t_n for each segment n, the k final detection functions can be obtained as $x_k(\tau)$ with $\tau = t_1, \ldots, t_n$ and $x_k(t_n) = X_k(n)$.

The last step which remains to be explained is the conversion of $x_k(\tau)$ into binary landmark functions $\Lambda_k(t)$.

Merging and binary landmarks The classification output of each segment provides a prediction score for each broad phonetic class and one non-speech class. Due to the speech recognition task used in the experimental evaluation of this thesis (see Appendix A.2), the non-speech models are associated with very heterogeneous spectral content, including silence, breath intakes and filler symbols that comprise for example environment noises or radio jingles. Thus, the non-speech class does not fit into the articulatory



(a) segment-based

(b) frame-based

Figure 6.3.: Segment and frame-based acoustic observations used to train classifier $F_k(\mathbf{x}_n)$ in the segment-based case and $f_k(\mathbf{y}'_t)$ in the frame-based case.

model used for speech segmentation. To avoid the insertion of too many non-speech landmarks due to this heterogeneous acoustic content, segments that are classified as non-speech are removed from the landmark detection process. If the non-speech class is not the highest scoring class, it is discarded from the prediction scores and landmark extraction relies effectively on the $k = 5$ remaining BPCs.

Extracting $\Lambda_k(t)$ from $x_k(\tau)$ relies on an additional merging parameter $\delta \in [0, 1]$ that corresponds to a threshold guaranteeing that the accumulated probabilities of active landmarks $\Lambda_k(t) = 1$ at frame t exceed δ with $\sum_k \Lambda_k(t) x_k(t) > \delta$, which is described in Algorithm 6.2. Thus, one can study the balance between activating few phones at t for low values of δ, which might produce several erroneous landmarks but provide precise knowledge for the ASR system, and activating many phones at landmarks $t \in \tau$ for high values of δ.

For the following experiments, this extraction step is done using the proposed segment-based landmarks $x_k(\tau_k)$ to obtain $\Lambda_k(t)$ and using the frame-based classifier output $x_k^{(\mathbf{y})}(t)$ to obtain continuous binary predictions $\Lambda_k^{(\mathbf{y})}(t)$.

Algorithm 6.2 Extraction of landmarks from detection functions.

```
Input: $x_k(\tau)$, $\delta$

1. Initialize $\Lambda_k(t)$ with $\Lambda_k(t) = 0$ for every $k$ and $t$
2. For each $t \in \tau$:
   a) Order the phonetic classes $k$ according to their scores $x_k(t)$ from highest to
      lowest, with $\hat{k}_{0,t}$ corresponding to the class with the maximum prediction
      score and $\hat{k}_{k,t}$ to the class with the lowest score
   b) Set $\Lambda_{\hat{k}_{0,t}}(t) = 1$
   c) for $i \in \left[\hat{k}_{1,t}, \hat{k}_{k,t}\right]$:
        $\Lambda_i(t) = 1$ if $\sum_k \Lambda_k(t)\, x_k(t) < \delta$

Output: $\Lambda_k(t)$
```

6.3. Experiments

The main argument for the use of landmarks as additional acoustic models is the assumption that landmarks can make use of acoustic observations that are more invariant than standard frame-based observations and thus provide more accurate predictions about speech classes. Yet, landmarks have the drawback to be less frequent than regular frame-based acoustic observations. Thus, there is a trade-off between inserting few, but precise landmarks or inserting more, but less precise frame-based acoustic observations. To study this relation, it is necessary to compare two different ways to drive the decoding according to Equation 6.3. The first way is the use of phonetic landmarks $\Lambda_k(t)$ and the second is the use of regular frame-based phonetic classifications $\Lambda_k^{(\mathbf{y})}(t)$. First, these two types of landmarks are compared by evaluating the classification and detection performance before comparing the speech recognition performance.

6.3.1. Classification and detection

Phonetic classification Table 6.3 displays several segment-based classification results on the development set obtained by evaluating $X_k(n)$, with each classifier using different segment-based observation vectors with 10 bagging iterations and 30% of the whole training instances as randomly chosen bootstrap samples. The classifiers trained differ in the amount of acoustic features used per frame, varying from using all 62 selected features to using only the 35 highest ranked features and they also differ in the fact whether segment-based observations of neighboring segments are included in the trained observation vector or not. Since feature selection selects features based on correlation instead of comparing the actual classification performance of the feature set, MFCCs

normalization	acoustic features (per frame)	context	accuracy
not normalized	35	yes	69.7
mean and variance	35	yes	71.7
quantiles ($q = 128$)	35	yes	71.8
quantiles ($q = 128$)	62	yes	73.4
quantiles ($q = 128$)	62	no	72.3
quantiles ($q = 128$)	62 + all 39 MFCCs	no	72.5

Table 6.3.: Classification results of segments $X_k(n)$ on the development set using 1/4 of the segmented ESTER 2 training corpus as training data. A segment was considered to be correct if the force-aligned reference utterance $q_{\mathcal{P}}(t)$ matches with the highest scoring class $X_k(n)$ at the temporal center of each segment.

might still be able to span an observation space that allows the learning of more discriminative decision boundaries. Therefore, the classification experiments are additionally carried out adding the whole set of 39 MFCCs to the acoustic features.

First, it can be seen that quantile normalization does not provide a significant improvement over mean and variance normalization. Using all 62 selected features per frame increased the classification accuracy 1.6%, compared to using only 35 features. Equally, concatenating each segment-based observations with its neighbors improves classification 1.1%. The limited improvement (0.2%) when the full set of MFCCs is added to the acoustic features in the context-free training confirms the result from feature selection, i.e., the full spectral observations do not provide significantly more information than the chosen acoustic parameters.

Frame and segment-based detection Both outputs $\{x_k(\tau)\}_k$ and $\left\{x_k^{(\mathbf{y})}(t)\right\}_k$ are evaluated with regards to their detection performance using conventional evaluation metrics precision and recall, derived from correctly predicted instances, incorrectly predicted instances and missed instances. To allow a fair comparison between landmark and frame-based detection functions $\{x_k(\tau)\}_k$ and $\left\{x_k^{(\mathbf{y})}(t)\right\}_k$, the landmarks are evaluated using the speech units obtained from of force-aligned reference utterances as reference instances. The detection task thus consists in correctly predicting the phonetic class of each aligned speech unit using frame-based phonetic class predictions derived from $\{x_k(\tau)\}_k$ and $\left\{x_k^{(\mathbf{y})}(t)\right\}_k$. Ideally, each aligned subword unit is correctly classified by at least one frame inside its boundaries. After mapping each aligned phone onto its broad phonetic class, the following evaluation metrics are calculated given binary landmarks:

- Correct prediction: A speech unit is considered as correctly predicted, as long as all frame-based predictions inside the boundaries of the speech unit predict the correct phonetic class of the unit

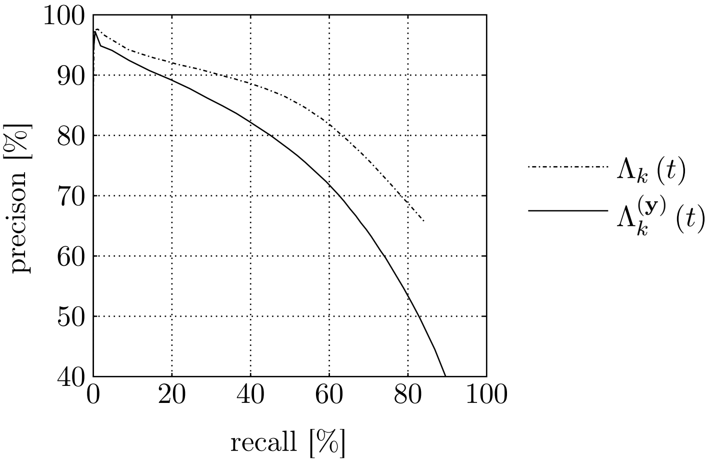


Figure 6.4.: Precision-recall curves for phonetic landmarks ($\Lambda_k\left(t, \delta^{(PR)}\right)$) and regular frame-based phonetic predictions ($\Lambda_k^{(\mathbf{y})}\left(t, \delta^{(PR)}\right)$), obtained by thresholding $\{x_k(\tau)\}_k$ and $\left\{x_k^{(\mathbf{y})}(t)\right\}_k$.

- Incorrect prediction: As soon as there is one falsely predicted frame t inside the speech unit boundaries, the unit is considered as an incorrect prediction
- Missed instance: If there is no active landmark inside the speech unit at all, the unit is considered as a missed instance

It should be noted that these three criteria are calculated globally using all phonetic classes and not for every phonetic class individually. To obtain precision-recall curves, both detection functions $\left\{x_k^{(\mathbf{y})}(t)\right\}_k$ and $\{x_k(\tau)\}_k$ are thresholded according to a precision-recall threshold $\delta^{(PR)}$ to activate binary landmarks in the segment case according to

$$\Lambda_{\hat{k}}\left(t, \delta^{(PR)}\right) = \begin{cases} 1 & \text{if } x_{\hat{k}}(t) > \delta^{(PR)} \text{ and } t \in \tau \\ 0 & \text{else} \end{cases} \tag{6.6}$$

with $\hat{k}$ corresponding to the highest scoring class at frame t. The activation in the frame-based case consequently corresponds to

$$\Lambda_{\hat{k}}^{(\mathbf{y})}\left(t, \delta^{(PR)}\right) = \begin{cases} 1 & \text{if } x_{\hat{k}}^{(\mathbf{y})}(t) > \delta^{(PR)} \\ 0 & \text{else} \end{cases} \tag{6.7}$$

This evaluation method allows to verify whether using the proposed segmentation and classification approach for landmark detection gives accurate phonetic predictions that result in more precise binary landmarks which cannot be achieved by simply choosing a suitable threshold on conventional frame-based prediction scores.

Figure 6.4 displays the precision-recall curves of the landmarks obtained from segment-based detection $x_k(\tau)$ and frame-based predictions $x_k^{(\mathbf{y})}(t)$. If the curves would be equal,

	$\bar{\Lambda}$		incorrect speech units [%]	
δ	$\Lambda_k(t)$	$\Lambda_k^{(\mathbf{y})}(t)$	$\Lambda_k(t)$	$\Lambda_k^{(\mathbf{y})}(t)$
-	8	8	31	85
0.8	14	16	13	45
0.95	22	24	4	17

Table 6.4.: Average number of activated phones per landmark ($\bar{\Lambda}$) and percentage of incorrect detected speech units using binary landmarks $\Lambda_k(t)$ and $\Lambda_k^{(\mathbf{y})}(t)$.

one could obtain the precision of the proposed segment-based landmark-detection algorithm by simply introducing a threshold into the frame-based classifications. Yet, it can be seen that at a recall above 40%, the precision of the landmarks is outperforming prediction of frames by roughly 10%. While no additional processing of the frame-based detection function $x_k^{(\mathbf{y})}(t)$, like smoothing, has been tested to improve frame-based landmarks, the comparison between the two knowledge sources shows that the segment-based framework allows more precise predictions about phonetic classes than frame-based observations, since it avoids to model portions of the signal with high acoustic ambiguity. This validates the initial idea of using a segment-based detection framework for improved prediction of phonetic classes at certain time instances.

Influence of merging parameter Table 6.4 displays how the threshold parameter δ (see Algorithm 6.2) influences the number of detection errors and the number of phones $\bar{\Lambda}$ activated at each frame, after extracting binary landmarks according to Algorithm 6.2. $\bar{\Lambda}$ measures the average number of phones activated at time instances $t \in \tau$, formally calculated as

$$\bar{\Lambda}\left(\{\Lambda_k(t)\}_k\right) = \frac{1}{|\tau|}\sum_{t\in\tau}\sum_k \Lambda_k(t)\,|\mathcal{S}_k| \tag{6.8}$$

with $|\cdot|$ corresponding to the cardinality of a set. $\bar{\Lambda}\left(\left\{\Lambda_k^{(\mathbf{y})}(t)\right\}_k\right)$ is calculated accordingly, except that τ in Equation 6.8 simply corresponds to all frames of the utterance. Detection errors correspond to speech units that have at least one frame t inside the boundaries of the speech unit, that attributes the correct class k with $\Lambda_k(t) = 0$, while activating another phonetic class at the same frame t.

Table 6.4 shows that while increasing δ leads to less speech units that are incorrectly detected, this comes at the cost of many phonetic classes k being activated at time instances $t \in \tau$, so that modifying the decoding according to these landmarks probably leads to few complementary information about the best path. It can be seen that there is a huge degree of freedom between choosing to insert landmarks with very low error rates (as low as 4% detection errors) and many activated phones (22 out of 39 possible phones) or landmarks with as few as eight activated phones that result in 31% detection errors.

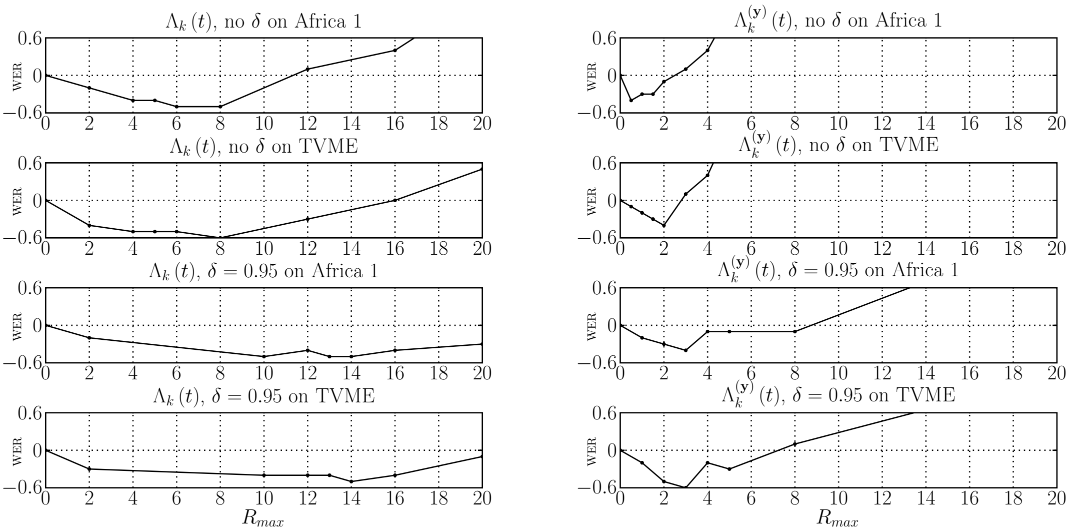


Figure 6.5.: Comparison of absolute WER improvement as a function of R_{max} for different values of δ on two radio stations (TVME and Africa 1) of the development set using either $\Lambda_k(t)$ or $\Lambda_k^{(\mathbf{y})}(t)$ to drive the decoding.

6.3.2. Speech recognition

Appendix A.2 gives a detailed explanation of the ESTER 2 broadcast news speech corpora that is used as training, development and testing data in the experiments of this thesis. The baseline speech recognizer corresponds to the triphone system discussed in Appendix A.3. Speech recognition experiments make use of landmarks in the first pass, while WERs reported correspond to the WERs obtained after the second pass. The landmarks are trained on narrow and wideband speech using 1/4 of the ESTER 2 training data with a nominal attribute for bandwidth as additional feature. The acoustic observation vector corresponds to the best setting in Table 6.3.

With the merging threshold δ and the weighting factor R_{max}, there are two parameters that have to be optimized for both, segment-based landmarks $\Lambda_k(t)$, as well as the decoding driven by $\Lambda_k^{(\mathbf{y})}(t)$, which is done by a grid-search on the development set. Thus, the optimal R_{max} is determined as the value that achieves the minimum WER on the development set. Figure 6.5 displays the improvement in WER as a function of R_{max} for $\Lambda_k(t)$ and $\Lambda_k^{(\mathbf{y})}(t)$ on two radio stations of the development set. The left column corresponds to recognition driven by $\Lambda_k(t)$, the recognition with frame-based scores $\Lambda_k^{(\mathbf{y})}(t)$ is on the right. The class merging scheme was not employed in the upper two rows, while $\delta = 0.95$ was used in the lower two rows. There are two observations worth mentioning from studying this figure:

- The improvements in WER are comparable for all different configurations, the only difference is the optimal R_{max}. Thus, as a rule of thumb, introducing more errors into the decoding requires a lower value for R_{max}. Yet summing up $\sum_k \Lambda_k(t) R_{max}$ results into a similar score for the best hypotheses for all settings. Errors can be

	baseline		$\Lambda_k^{(\mathbf{y})}(t)$		$\Lambda_k(t)$	
δ	-		0.8		0.6	
R_{max}	-		2		8	
broadcast	dev	test	dev	test	dev	test
Inter	21.9	18.7	21.7	18.4	21.6	18.6
RFI	-	17.6	-	17.3	-	17.2
Africa 1	45.1	31.5	44.6	30.7	44.7	30.8
TVME	30.1	24.1	29.6	24.2	29.6	24.4
all	-	23.5	-	23.1	-	23.2

Table 6.5.: ASR results on the development and test set using landmark-based ($\Lambda_k(t)$) and frame-based ($\Lambda_k^{(\mathbf{y})}(t)$) guiding of the decoding for different broadcast news shows.

increased by decreasing δ or by changing from landmark-based decoding with $\Lambda_k(t)$ to frame-based decoding using $\Lambda_k^{(\mathbf{y})}(t)$.

- The landmarks that provided the least amount of errors, i.e. , $\Lambda_k(t)$ with $\delta = 0.95$, are the least sensitive to the exact choice of R_{max}. Consequently, there is no difference in the optimal R_{max} for the Africa 1 broadcast shows and the optimal R_{max} of the TVME broadcast shows (left column and lower two rows in Figure 6.5) . In contrast to that, $\Lambda_k^{(\mathbf{y})}(t)$ differs significantly in the optimal R_{max} between the two broadcast shows (right column and upper two rows in Figure 6.5).

Results on the ESTER 2 test set using the optimal R_{max} and δ obtained on the development set are displayed in Table 6.5. The overall improvement in WER with respect to the 23.5% WER of the baseline of the test set was 0.3% using $\Lambda_k(t)$ and 0.4% for $\Lambda_k^{(\mathbf{y})}(t)$. Statistical significance of the WER improvement was tested using a Wilcoxon signed-rank test and it proved to be significant at the 5% level, for both cases. Radio Inter achieved the smallest gain on the development set with 0.2% ($\Lambda_k^{(\mathbf{y})}(t)$) and 0.3% ($\Lambda_k(t)$) but confirmed this small gain on the test set. RFI was not included in the development set, but gained 0.3% and 0.4%. Africa 1 performed well and improved the WER 0.8% and 0.7% on the test set. While TVME gained 0.5% for both approaches on the development set, this was not confirmed on the test set. The difference between using $\Lambda_k^{(\mathbf{y})}(t)$ for the Viterbi decoding and using $\Lambda_k(t)$ is not significant, but it has to be emphasized that simple frame-based phonetic class predictions improve speech recognition equally well as the presented landmark detection framework which is further discussed in the following conclusions.

6.4. Conclusions

The chapter changed the integration of phonetic landmarks concerning two aspects compared to the acoustic pruning in the previous chapter. First, the integration was changed towards a combination approach, combining weighted binary landmarks with the emission probabilities of the ASR system. Second, a landmark detection approach that made careful use of phonetic knowledge in a segmentation and classification approach was designed for external landmark detection and applied inside the baseline ASR system using the proposed decoding. While the speech recognition results showed an improvement compared to the baseline, there was no advantage of using landmarks over regular continuous frame-based BPC classifiers.

There are several possibilities to improve some parts of the proposed landmark detection approach, including extending the acoustic observations, possibly towards incorporating information at segment boundaries, using more advanced statistical classifiers or extending the segmentation to a multi-level representation with different degrees of sensitivity to spectral changes at each level.

It can be concluded that while segment-based phonetic landmarks seemed to provide advantages over regular continuous frame-based phonetic predictions in the sense that their predicted landmarks achieved a higher precision on a phonetic detection task, both approaches improve the baseline speech recognizer equally. Thus, in the end the products of $\Lambda_k\left(t\right)$ and R_{max}, respectively $\Lambda_k^{(\mathbf{y})}\left(t\right)$ and R_{max} led to a similar improvement of the obtained n-best list. While potential reasons for this result will be further discussed after the next chapter, one can already see that, while extending the acoustic observations allowed more precise predictions at frames $t \in \tau$, the use of a shared framework to obtain the k detection functions $\left\{x_k\left(\tau\right)\right\}_k$ did not allow for detailed phonetic modeling, apart from using phonetically motivated segmentation and acoustic observations, which might not lead to truly complementary acoustic information of landmarks over continuous frame-based modeling.

While the next chapter will not directly attempt to improve the presented approach, it will present a second landmark-detection approach that builds upon segment-based acoustic observations, which showed to provide more accurate phonetic predictions in this chapter, but will reduce the amount of phonetically motivated heuristics and focus on improving landmark-driven ASR by more advanced statistical modeling of phonetic classifiers.

7. Landmark detection using segment-based classifiers

While the landmark detection algorithm presented in this chapter is not a direct follow-up on the approach presented in the previous chapter, it is also based on predicting phonetic classes for time-variable speech segments, but comes with two major differences. First, extracting the segment-based observation vector for a speech segment relies on the use of dynamic programming to obtain spectral homogeneous subsegments, which are used to warp the dynamic content of a speech segment onto a fixed-dimensional observation vector. The speech segments are no longer obtained by data-driven segmentation, but standard speech transcriptions obtained by aligning HMM-based phone models. Second, segment-based phonetic classifiers are embedded in a classical «peak-picking» detection algorithm, which uses trained classifiers to obtain continuous detection profiles, which are further processed to obtain potential landmarks as local maxima of the detection profiles. These landmarks are converted into binary landmark sequences and integrated into the search by weighted combination, as proposed in the previous chapter.

7.1. Motivation and overview

One of the simplest, yet widely used approaches to landmark detection is to train k frame-based classifiers, or one multi-class classifier, to attach every frame of the speech utterance with a prediction score to obtain k continuous detection profiles $x_k(t)$, from which landmarks are extracted as the local maxima (e.g., [EWPJD07, JN08]). This method implies that the nuclei of speech events match well the general model trained for each class, so that speech events can be identified as local maxima in the obtained detection function. The amplitude of a local maximum is a measure of quality of the detected event, since the higher the amplitude $x_k(t)$ at t, the better the match between the acoustic observation vector at t and the trained classifier.

This assumption has two major limitations. First, frame-based models rely to a certain extend on noisy training observations, since some frames will be corrupted due to the variability of speech, especially coarticulation, and frames belonging to long speech units will be overrepresented in the training data. Second, these methods propagate many of the disadvantages of the «beads-on-a-string» model of speech into the detection of speech events, since classifiers are trained to predict the speech label of frame-based acoustic observations, which can only capture fractions of the acoustic correlates of articulatory gestures, as it has been discussed in Chapter 3. Associating local maxima of $x_k(t)$ with the most salient points of articulatory gestures is therefore a rather crude heuristic, which is hard to justify by phonetic studies. This chapter therefore proposes

Algorithm 7.1 Extraction of a fixed-dimensional observation vector for time-variable segments

Input: $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_t$, $\{(s_n, e_n)\}_n$

1. Divide the speech signal $\mathbf{Y}$ into n segments $\left\{\mathbf{Y}_{s_n}^{e_n}\right\}_n$.
2. Determine two segment boundaries $b_2^{(n)}$ and $b_3^{(n)}$, with $b_1^{(n)} = s_n$ and $b_4^{(n)} = e_n + 1$ for each phone segment to divide $\mathbf{Y}_{s_n}^{e_n} = \mathbf{y}_{s_n}, \ldots \mathbf{y}_{e_n}$ into $i = 3$ subsegments $\left\{\mathbf{Y}_{b_i^{(n)}}^{b_{i+1}^{(n)}-1}\right\}_i$.
3. Extract a fixed-dimensional observation vector for each subsegment and concatenate these vectors to a final observation vector $\mathbf{x}_{(s_n,e_n)}$ for each segment n.

Output: $\left\{\mathbf{x}_{(s_n,e_n)}\right\}_n$

the use of segment-based acoustic observations for landmark detection, since segments are intuitively a better model for time-variable speech units, as it has been proposed for ASR in [Ost99, Gla03] and already discussed in Chapter 3. The goal is to train classifiers that extract acoustic observations from temporal variable subword units and apply these classifiers sequentially on speech utterances to obtain improved detection profiles for landmark extraction.

7.2. Algorithm

The main idea of the algorithm presented in this chapter is to train a classifier that predicts the phonetic class of a given segment. In this chapter, a segment always corresponds to a potential subword speech unit of the baseline ASR system. Since the classifier is trained on the standard force-aligned phone segments, one has to provide unlabeled segments during prediction. To avoid the use of a priori segmentation of speech utterances, the classifier is applied on a fixed amount of overlapping segments to obtain a classical detection function.

The two main components of the algorithm are displayed in Algorithm 7.1 and Algorithm 7.2. Algorithm 7.1 describes the extraction of a fixed-dimensional observation vector for time-variable speech units, given a sequence or collection of speech segments, with the segment boundaries for example being obtained by force alignment of phone models. The observation vector for a segment is obtained by dividing the segment into three subsegments, extracting one vector for each subsegment and concatenating all three vectors to one final observation vector (see Section 7.2.1). The obtained observation vector is used to train a classifier that is applied in Algorithm 7.2 to obtain a frame-based detection function $x_k(t)$ for k phonetic classes given an a priori collection of segments $\{(s_n, e_n)\}_n$. $x_k(t)$ is further processed by a classical peak-picking algorithm to obtain k binary landmark indicators $\Lambda_k(t)$ (see Section 7.2.2).

Algorithm 7.2 Obtaining a frame-based detection profile using a segment-based classifier

```
Input: $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_t$, $F(\mathbf{x}_{(s,e)})$, $\{(s_n, e_n)\}_n$
  1. Use the collection of $n$ (potentially overlapping) segments $\{(s_n, e_n)\}_n$ indicating
     boundaries of potential speech events to extract the observation vector according
     to Algorithm 7.1 for each segment to obtain $n$ observation vectors $\left\{\mathbf{x}_{(s_n,e_n)}\right\}_n$.
  2. Attribute a score to each segment using $\mathbf{x}_{(s_n,e_n)} \in \left\{\mathbf{x}_{(s_n,e_n)}\right\}_n$ for each phonetic
     class $k$ using classifier $F_k(\mathbf{x}_{(s,e)})$.
  3. To obtain a detection function $x_k(t)$ for each speech class $k$, search for each frame
     $t$ the segment that maximizes the detection score at this frame.
         $x_k(t) = \max_n F_k(\mathbf{x}_{(s_n,e_n)});\ s_n \leq t \leq e_n$
  4. Extract binary landmarks $\Lambda_k(t)$ according to the peak-picking method presented
     in Algorithm 7.4.
Output: $\Lambda_k(t)$
```

7.2.1. Extraction of fixed-dimensional observation vectors from time-variable segments

To obtain a fixed-dimensional observation vector for a speech unit segment, each speech unit is divided into three subsegments, with each subsegment ideally corresponding to a spectral homogeneous section of the speech unit. The subsegments are limited to three, since subword units used in this thesis correspond to phones that can be divided into the left and right context and the phone nucleus. The goal is to extract one acoustic observation vector for each subsegment, to obtain a compressed version of the dynamic spectral information of the whole speech segment by concatenating all three vectors. To obtain optimal subsegment boundaries, the following algorithm determines intra-phone boundaries according to an intra-segment distortion measure, like it has been proposed in [SRS02] for segmenting speech utterances into a sequence of citation-phonemic speech units.

Subsegmentation Given the parametrized frame-based representation of a speech unit $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_n$, in this chapter corresponding to a sequence of MFCC vectors, with n corresponding to the length in frames of the speech unit, the segmentation algorithm searches for the borders b_2 and b_3 segmenting the unit into $i = 3$ subsegments. This results in the subsegments $(\mathbf{y}_{b_1}, \ldots, \mathbf{y}_{b_2-1})$, $(\mathbf{y}_{b_2}, \ldots, \mathbf{y}_{b_3-1})$ and $(\mathbf{y}_{b_3}, \ldots, \mathbf{y}_n)$, with $b_1 = 1$. Thus, the minimum length of a speech unit that can be segmented corresponds to three frames. The distortion criterion measures the intra-segment distortion of each subsegment as the accumulation of distances from each frame inside the subsegment to its segment-centroid μ_i. Using the euclidean distance between frames as a distance

Algorithm 7.3 Dynamic programming algorithm for subsegmentation of speech units.

Input: $\mathbf{Y}_{(s,e)} = \mathbf{y}_s, \ldots, \mathbf{y}_e = \mathbf{y}_1, \ldots, \mathbf{y}_n$

1. Initialize each element $\mathbf{D}_{a,b}$ of a $n+1$ dimensional distance matrix $\mathbf{D}$ with infinity:

$$\mathbf{D}_{a,b} = \infty$$

2. Calculate the values $\mathbf{D}_{a,b}$ with $b \geq a$ above the diagonal of $\mathbf{D}$:

$$\mathbf{D}_{a,b} = \sum_{n=a}^{b-1} \|\mathbf{y}_n - \mu_{(a,b)}\|$$

3. Use dynamic programming to find the frames corresponding to the borders $i = 2$ and $i = 3$ by keeping track of the best path by calculating

$$q_i\left(b\right) = \min_a q_{i-1}\left(a\right) + \mathbf{D}_{a,b}$$

with $q_1\left(1\right) = 0$ and $q_1\left(n > 1\right) = \infty$. The optimal path corresponding to the segment boundaries b_1, b_2, b_3 and b_4 is obtained using back-pointers and backtracking from $q_4\left(n+1\right)$.

Output: b_1, b_2, b_3, b_4

measure, the search for the borders b_2 and b_3 given $b_1 = 1$ and $b_4 = n+1$, is carried out by minimizing the segment distortion according to:

$$b_i = \arg\min_{b_i} \sum_{i=1}^{3} \sum_{n=b_i}^{b_{i+1}-1} \|\mathbf{y}_n - \mu_i\|\,;\; b_1 = 1, b_4 = n+1. \tag{7.1}$$

Finding the optimal segmentation corresponds to a shortest-path problem, which can be solved for each segment using dynamic programming according to Algorithm 7.3.

Obtaining the final observation vector and classification Each subsegment is reduced to its maximum-likelihood estimation μ_i, corresponding to the centroid of each subsegment, and concatenating these three vectors results in an observation vector $\mathbf{x}'_{(s,e)} = [\mu_1, \mu_2, \mu_3]$ for each speech unit, spanning from frame-index s to frame-index e, with s corresponding to the first frame of the speech unit and e corresponding to the last frame of the speech unit. As in Chapter 6, the overall length of the segment $l_{(s,e)} = e - s$ and the length of each subsegment in frames $l_i = b_{i+1} - b_i$ are added as additional attributes to each $\mathbf{x}'_{(s,e)}$. The three intra-segment distortion measures $d_i = \sum_{t=b_i}^{b_{i+1}-1} \|\mathbf{y}_t - \mu_i\|$ are additionally added to provide information about the spectral homogeneity of the subsegment, so that the final observation vector corresponds

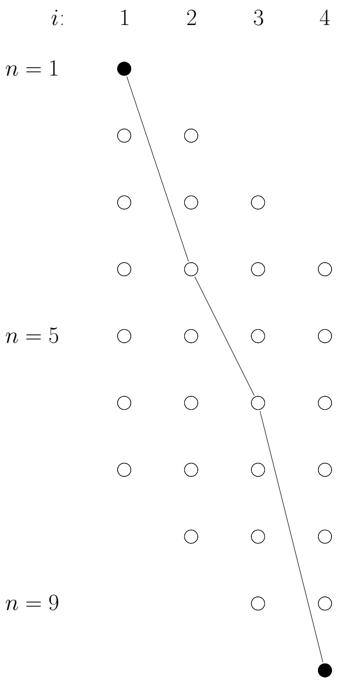


Figure 7.1.: Example of the backtracking trellis of a speech unit with a length of 9 frames. The arc between two nodes corresponds to the distance measure $\mathbf{D}_{a,b}$ and the example shows the best path tracked back from the final node $n + 1$ resulting in subsegments from frame 1 to frame 3 , frame 4 to frame 5 and frame 6 to frame 9.

to:

$$\mathbf{x}_{(s,e)} = \left[\mathbf{x}'_{(s,e)}, \left\{d_{i,(s,e)}\right\}_i, \left\{l_{i,(s,e)}\right\}_i, l_{(s,e)}\right]$$

This observation vector is extracted for all subword units of the training utterances to train a classifier $F_k(\mathbf{x}_{(s,e)})$ for k phonetic classes using decision stumps in connection with the AdaBoost.MH [FS95, FSA99, SS99] boosting algorithm as statistical classifier. Decision stumps correspond to shallow decision trees. Thus, the classifier is still based on binary decision rules, like it has been proposed in the previous chapter, but since the acoustic observations $\mathbf{y}_t$ do not correspond to acoustic parameters but MFCCs, the goal is not to obtain phonetic rules similar to acoustic-phonetic ASR, but to obtain accurate statistical decision boundaries and boosting has shown to be one of the most accurate ensemble methods for classification [BK99].

The trained classifier consists of i weak binary learners $h_i(\mathbf{x})$ and associated weights α_i that can be linearly combined to obtain a real valued prediction score $f(\mathbf{x}) = \sum_i \alpha_i h_i$. AdaBoost.MH [SS99] extends the binary problem to the multi-class case by splitting the multi-class problem into several binary problems, so that the final classifier $f(\mathbf{x})$ effectively provides a prediction score $f_k(\mathbf{x})$ for k classes. To avoid training a single model on the whole ESTER 2 training corpus, which would be computational expensive

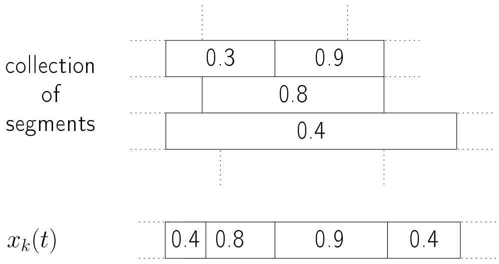


Figure 7.2.: Simplified example for obtaining a detection function $x_k\left(t\right)$ from a collection of segment-based scores.

given the huge amount of training instances, the training set is divided into M non-overlapping partitions and an individual classifier $f^{(m)}\left(\mathbf{x}\right)$ is trained on each partition (see [CMH$^+$03]). The final classification score $F_k(\mathbf{x})$ is obtained by averaging the output scores of the full committee of classifiers :

$$F_k(\mathbf{x}) = \frac{\sum_{m=1}^{M} f_k^{(m)}(\mathbf{x})}{M}. \tag{7.2}$$

The following section describes the use of a segment-based classifier $F_k(\mathbf{x}_{(s,e)})$ to determine a detection function $x_k\left(t\right)$ for each phonetic class k, from which the final sequence of landmarks $x_k\left(\tau\right)$ is extracted.

7.2.2. Obtaining frame-based detection functions

To solve the issue of extracting a frame-based detection function $x_k\left(t\right)$ using a segment-based classifier, the following method assumes that the optimal value for the detection profile $x_k\left(t\right)$ at frame t for phonetic unit k is obtained by searching the segment that maximizes the score of the phonetic class k at frame t. A frame t is thus never considered in isolation but rather in context of potential speech units that include frame t. While this method seems very intuitive, it should be noted that the k scores at each frame t do not correspond to strict probability estimates any more, since the score of each class k at frame t is not obtained by comparing identical acoustic observations.

From classification to detection Since the trained classifier is only able to classify acoustic observations extracted from given segments, the speech signal has to be a priori segmented into n overlapping potential speech segments and corresponding observation vectors $\left\{\mathbf{x}_{(s_n,e_n)}\right\}_n$, with s_n and e_n corresponding to the first and last frame of the n-th segment. A simple prior segmentation is a phone lattice, obtained from the n-best output of a phone-based decoding of the utterance, yet this method is not used in this algorithm to avoid the usage of additional models for obtaining the phone lattice and to avoid missing out on potential speech events due to errors in the lattice. Therefore, the algorithm makes use of an exhaustive prior segmentation as collection of strongly overlapping segments $\left\{(s_n, e_n)\right\}_n$, by predicting the observations extracted from all possible segments inside the speech signal, up to a maximum segment length of 300ms. This

Algorithm 7.4 Peak-picking algorithm for one speech utterance.

```
Input: $x_k\left(t\right)$

1. Collect the ensemble of local maxima $\tau_k$ for each detection function $x_k\left(t\right)$ and
   create the union of all local maxima $\tau = \bigcup_k \tau_k$ to obtain potential landmarks.

2. Determine a threshold $\delta(\tau)$ for each local maximum $t \in \tau$. $\delta(t)$ corresponds to
   $x_k(t)$ if there is only one class $k$ for which $t \in \tau_k$. Else, if several classes share a
   local maximum, the class $k$ with the highest amplitude $x_k(t)$ and $t \in \tau_k$ provides
   $\delta(t)$.
   for $t \in \tau$:
      for all $i$ for which $t \in \tau_i$:
         $\delta(t) = \max_i x_i(t)$

3. Initialize the phonetic landmarks of each class $k$ with $\Lambda_k\left(t\right) = 0$ for all $t$.

4. Activate a phonetic class at each local maxima if the amplitude of the detection
   function is above the defined threshold:
   for $t \in \tau$:
      foreach $k$:
         if $x_k(t) \geq \delta\left(t\right)$:
            $\Lambda_k\left(t\right) = 1$

Output: $\Lambda_k\left(t\right)$
```

corresponds to 30 segments per frame, not counting frames at utterance endings. The two computationally most costly parts that have to be considered when such a large amount of segments is used, is the calculation of the distance matrix $\mathbf{D}$ (see Algorithm 7.3) for obtaining the observation vector and the prediction of the score for each segment. The computational load due to the calculation of the distance matrix can be reduced for each utterance by sharing a common pool for all the distances $\mathbf{D}_{a,b}$ that are computed and shared across overlapping segments.

The optimal value of the detection function $x_k\left(t\right)$ at each frame t, given n overlapping segments and corresponding observation vectors $\left\{\mathbf{x}_{(s_n,e_n)}\right\}_n$, is calculated as follows,

$$x_k\left(t\right) = \max_n F_k(\mathbf{x}_{(s_n,e_n)});\ s_n \leq t \leq e_n \tag{7.3}$$

with $F_k(\mathbf{x})$ being the prediction score for the phonetic class k, given observation $\mathbf{x}$. A simplified example for a collection of segments with attached prediction scores and the resulting detection function is displayed in Figure 5.2.

Peak-picking with dynamic threshold While the obtained continuous detection function should be less noisy, compared to standard frame-based detection functions, there still will be many low scores $x_k\left(t\right)$ at local maxima that have to be filtered out before trained on 1/12 of training set

#	classifier	accuracy
1	$FRAME$ (depth=2)	55.7
2	$SEG_{1/12}$ (depth=1)	57.8
3	$SEG_{1/12}$ (depth=2)	63.0
4	$SEG_{1/12}$ (depth=3)	62.8

trained on full training set

#	classifier	accuracy
5	HMMs (64 Gaussians)	65.9
6	$SEG_{12/12}$ (depth=2)	67.6

Table 7.1.: Phoneme classification results on the ESTER 2 development set using different acoustic observations. Classification was performed by predicting the phoneme label of the force-aligned reference utterances. «depth» corresponds to the depth of the decision stumps used as weak learners during boosting.

converting $x_k(t)$ into binary landmarks $\Lambda_k(t)$. The peak-picking algorithm 7.4 proposed in this section consists of two steps. First, local maxima τ_k are extracted from each detection profile $x_k(t)$ and used to determine a threshold $\delta(t)$ at all local maxima $t \in \tau$ with $\tau = \bigcup_k \tau_k$. If several classes k share a local maximum at t, only the highest amplitude is used as the threshold $\delta(t)$. Binary phonetic landmarks $\Lambda_k(t)$ are then obtained by activating landmarks with $\Lambda_k(t) = 1$, if a local maxima $x_k(\tau_k) \geq \delta(t)$. Thus, each local maximum $t \in \tau$ will propagate into a binary landmark at frame t, yet several phonetic classes k will be activated at frames t with $\Lambda_k(t) = 1$, in case the local maximum corresponds to a low value to prevent false alarms.

7.3. Experiments

The experimental evaluation of the proposed method involves a phoneme classification task, to study the prediction power of the proposed segment-based observation vector and a short discussion about the obtained landmarks, before landmark-driven speech recognition is evaluated on the same recognition task as the experiments in Chapter 6.

Phoneme classification experiment The presented algorithm is not particularly adapted to phonetic classes, but compatible with any set of speech unit labels. Comparing the classification performance of the trained statistical classifiers on a phoneme classification task is therefore a valid way to examine the ability of the proposed observation vector to capture the fine acoustic distinctions between phonemes, rather than phonetic classes. The phoneme set corresponds to the phone inventory of the speech recognizer and classifiers are trained to predict the phoneme label given the phoneme boundaries obtained from reference alignments.

For the experiments, several ensembles of boosted decision stumps are trained using the acoustic observations obtained by applying Algorithm 7.1, given the aligned speech units. The following experiments contain prediction results from two segment-based classifiers, referred to as classifiers $SEG_{m/M}$, with m referring to the number of partitions of the full training set used for training, selected from the full set of M partitions. The baseline, against which the trained classifiers are compared, corresponds to monophone 3-state left-to-right HMMs with 64 diagonal-covariance Gaussian components per state, as described in Appendix A.3. HMMs are trained on the full training set using the same speech parametrization as the boosted ensembles, which are 39-dimensional MFCC vectors. Additionally to HMMs, the classifiers trained with the proposed observation vector are compared to the segment-based observation vector used in Chapter 6, i.e., the concatenated three frames at the center of each speech unit, equally trained using boosted decision stumps (referred to as classifier $FRAME$), for which the dimensionality of the observation vector is nearly identical.

For all experiments, the number of boosting rounds was limited to 3,000. Table 7.1 displays the classification accuracy on the development set as the percentage of correctly classified phonemes. Using the proposed observation vector for boosting (classifier $SEG_{1/12}$) increases the accuracy by 7.3% (using decision trees with depth=2) compared to classifying speech units based on three concatenated frames ($FRAME$). The improved performance of $SEG_{1/12}$ over $FRAME$ shows two things. First, it can be assumed that considering the spectral trajectory, as it is attempted in the proposed segment-based observation vector, captures more relevant information that helps to identify the speech unit than it is possible when only the acoustic content at the center of a speech unit is considered. Second, if $FRAME$ is used to obtain a frame-based detection function, the lacking ability to capture all relevant acoustic information and the resulting lack in precision is likely to propagate the classification errors into the detection step, by creating unreliable detection functions $x_k\left(t\right)$.

Comparing the classification performance to classification using HMMs shows that only by using the ensemble of classifiers $SEG_{12/12}$ classification performance exceeds HMMs by 1.7%. Since the three subsegments of the observation vector are likely to resemble the Gaussian mixtures of the three HMM states, the improvement is probably rather due to the statistical power of classifier combination of 12 classifiers, than due to additional acoustic information. This result might also indicate that the segment-based classifier $SEG_{12/12}$ is not likely to outperform more advanced acoustic models, like context-dependent HMMs on rescoring tasks.

Phonetic landmark extraction and detection performance To derive k phonetic scores for each segment (s_n, e_n) from the trained phone-based classifiers $FRAME$, $SEG_{1/12}$ and $SEG_{12/12}$, the score of the phonetic class k is approximated by the highest phone score of its associated set of phones $\mathcal{S}_k$, with

$$F_k(\mathbf{x}_{(s,e)}) = \max_{p \in \mathcal{S}_k} F_p^{(phone)}\left(\mathbf{x}_{(s,e)}\right),$$

	classifier		
	$FRAME$	$SEG_{1/12}$	$SEG_{12/12}$
number of landmarks	1,352k	574k	608k
detection errors [%]	67.7	20.2	18.0
missed speech units [%]	4.6	12.0	8.8
$\bar{\Lambda}$ [phones]	13	20	20

Table 7.2.: Landmark detection performance using $k = 7$ BPCs on the development set. Speech units corresponding to silence have been discarded from the evaluation.

so that $k = 7$ detection functions $x_k(t)$ are obtained for $k = 7$ BPCs vowels, nasals, approximants, fricatives (voiced and unvoiced) and plosives (voiced and unvoiced).

The detection functions for the classifiers $SEG_{1/12}$ and $SEG_{12/12}$ were obtained by using Algorithm 7.2. The detection function for classifier $FRAME$ was obtained by predicting each frame of the speech signal, using a context-window of one frame. All detection functions were used to extract binary landmarks $\Lambda_k(t)$ using the peak-picking Algorithm 7.4.

Table 7.2 contains information about the landmark detection accuracy using the same evaluation metrics as proposed in Chapter 6 for Table 6.4, i.e., average number of activated phones per landmark ($\bar{\Lambda}$) and detection errors (in percentage of misclassified speech units). Comparing the landmarks obtained by the frame-based detection function and the landmarks obtained by the proposed segment-based detection function has to be done with care, since the frame-based detection function has not been subject to smoothing or similar post-processing that could reduce the amount of noise present in the detection function. These noisy detection functions lead to twice as much local maxima, compared to the landmarks extracted with the proposed segment-based classifier. The merging step according to Section 7.2.2 activates few phones for the frame-based detection function, with 13 activated phones on average, compared to 20 in the segment-based case. These few activated phones in the frame-based case lead to more erroneous landmarks, which result in a huge amount of detection errors, compared to the rather moderate 20.2%, respectively 18%, errors for the segment-based cases.

Comparing these results with the landmarks extracted in the previous chapter, the landmarks obtained using the $SEG_{12/12}$ classifier have less missed speech units (about 8% less missed speech units), but the number of detection errors rises about 14% at comparable average phones activated per landmark (22 phones on average in the previous chapter, 20 phones on average using classifier $SEG_{12/12}$), which is partly due to using finer phonetic resolution by distinguishing between voiced and unvoiced consonants and slight differences in the utterances tested. Additionally, increasing the number of landmarks also increases the variance of the activated phones per landmark.

Speech recognition The setup of the speech recognition experiments is similar to the setup explained in Section 6.3, i.e., the k landmarks $\Lambda_k(t)$ are converted to a

broadcast	WER baseline	WER landmark-driven	best result from Chapter 6
Inter	18.7	18.6	18.4
RFI	17.6	17.3	17.3
Africa1	31.5	30.8	30.7
TVME	24.2	23.8	24.2
all	23.5	23.1	23.1

Table 7.3.: Speech recognition results using binary landmarks extracted from detection functions obtained from classifier $SEG_{12/12}$. The result for each broadcast is compared to the best result obtained from either $\Lambda_k^{(\mathbf{y})}(t)$ or $\Lambda_k(t)$ in Chapter 6 (see Table 6.5).

binary mask $\lambda_j(t)$ and integrated into the baseline using the weighted integration of binary landmarks according to Equation 6.3. The only exception was that landmarks are not trained or applied on narrowband speech. R_{max} as the only parameter to be estimated, was again obtained by a one-dimensional grid search on the development set. The landmarks used for speech recognition experiments were obtained using classifier $SEG_{12/12}$ and thus did not predict non-speech symbols at all, yet experiments showed no difference in WER in case non-speech symbols were included in the classification and prediction or not.

The results in Table 7.3 show an improvement for all four broadcast shows tested, with the improvement varying from 0.1 to 0.7 . The overall WER of the test set was 23.1%, which is the same improvement that was obtained by the best-performing set of landmarks from Chapter 6. Yet, it can be seen that while the landmarks employed in the previous chapter degraded the WER on one broadcast (TVME) the landmarks in this chapter achieved small but constant improvements on all broadcasts. Nevertheless, the global conclusion is unchanged from the conclusion in the last chapter, i.e., despite additional computational effort and a new landmark extraction method, the WER obtained by decoding with landmark-based methods and simple frame-based predictions is quasi identical.

7.4. Conclusions

This chapter provided a second alternative for the extraction of binary landmarks which are used for weighted combination with the emission probabilities of the baseline ASR system. The issue that was focused on was using a fixed-dimensional observation vector to describe variable-length segments and to extract landmarks by determining the local maxima of detection functions obtained by sequentially applying a segment-based classifier on the speech utterance. While the classification experiments indicated that the proposed observation vector provided better classification results than simple concatenation of frames, as used in Chapter 6, the weighted landmark-driven decoding did not significantly improve compared to the results obtained with the approach in Chapter 6.

Potential improvements of the described algorithm include to reduce the computational load by prior segmenting the utterance into a phone lattice, instead of using exhaustive segmentation. Since the subsegmentation is also computationally costly, one might want to compare subsegments obtained by dynamic programming to heuristic subsegmentation methods, which divide speech segments based on a fixed splitting rule.

The results obtained from this chapter allow several general conclusions about the two landmark detection frameworks presented. First, while the approach in Chapter 6 made use of phonetically motivated segmentation and acoustic features, this did not improve the WER compared to the segment-based acoustic observations used in this chapter, which used less phonetically motivated models. Using phonetic landmarks instead of standard frame-based predictions provided advantages concerning classification accuracy on classification and detection tasks, yet in the end the product of binary landmarks and R_{max} resulted in similar n-best hypotheses and improvement of the WER, compared to the baseline. This shows that the landmark detection front-end did provide few complementary information to the frame-based acoustic emission probabilities, which can be seen in the relatively small gain in WER achieved, and no complementary information compared to frame-based phonetic classification. The reason for this lack of complementary knowledge can be found in the use of a shared detection front-end, which allowed only homogeneous modeling of all broad phonetic classes. Indeed, it has to be admitted that besides changing phone alignments into phonetic labels and extended acoustic observations, the training method was very similar to the standard acoustic modeling. Obviously, the acoustic models can compensate their weak frame-based acoustic observations by the statistical power of context-dependent models and the language model, which makes the additional information by phonetic landmarks redundant.

Given this lack of complementary knowledge, an obvious improvement would be to switch to individual front-ends to optimize acoustic observations and post-processing individually for each phonetic class. Thus, at this point further extensions of shared front-ends, as the ones presented in the last two chapters, are left for future work and the last chapter attempts to attack the problem of landmark-driven ASR from a different perspective, by presenting a new general landmark integration framework, that can account for an arbitrary number of individual detection front-ends that provide heterogeneous and asynchronous landmark sequences.

8. A general framework for integrating heterogeneous and asynchronous landmarks into ASR

Since many approaches that attempt to integrate phonetic information into statistical ASR, including the ones presented in this thesis, encounter the issue of how to provide truly complementary information for the baseline ASR systems, it is worth to think about the fact that the problem of integrating phonetic knowledge into standard ASR might have to be solved from an «integration perspective», i.e., provide a framework that allows the integration of truly heterogeneous knowledge with a maximum possible degree of freedom, with the detailed development of individual detection front-ends being an ongoing part of research.

Such a framework might have two effects. First, it might encourage ongoing research for new phonetic detection front-ends that can pursue very individual detection strategies, if there are only minimal constraints on the required output type, and new detectors can immediately be applied inside landmark-driven ASR. Second, this framework might also be useful for external knowledge sources that go beyond phonetic class detection, for example the visual modality, which might be able to provide information about selected phones at certain time instances, if a landmark-like visual detection strategy is pursued.

To allow the ongoing integration of such heterogeneous knowledge, the knowledge sources that can be integrated with the framework presented in this chapter are defined as freely as possible according to Chapter 4, i.e., an arbitrary number of detection front-ends, with each front-end producing heterogeneous and asynchronous landmarks indicating the presence of an ensemble of phones. The framework attempts to integrate these detection functions into ASR by mapping each raw landmark sequence onto a log-likelihood score, representing the confidence that the landmark indicates the correct path. The obtained likelihoods are then discriminatively trained with the acoustic models and integrated into the Viterbi decoding using standard weighted linear combination.

8.1. Knowledge sources and combination

To integrate an arbitrary number of individual detection front-ends into standard ASR, the proposed knowledge integration framework consists of two parts (see also Figure 8.1). The first part maps each detection function $x_k(\tau_k)$ onto a log-likelihood score $\log s_k(t)$, with $\log s_k(t) \geq 0$, reflecting the likelihood of $q_{\mathcal{P}}(t)$ being member of $\mathcal{S}_k$, to account for

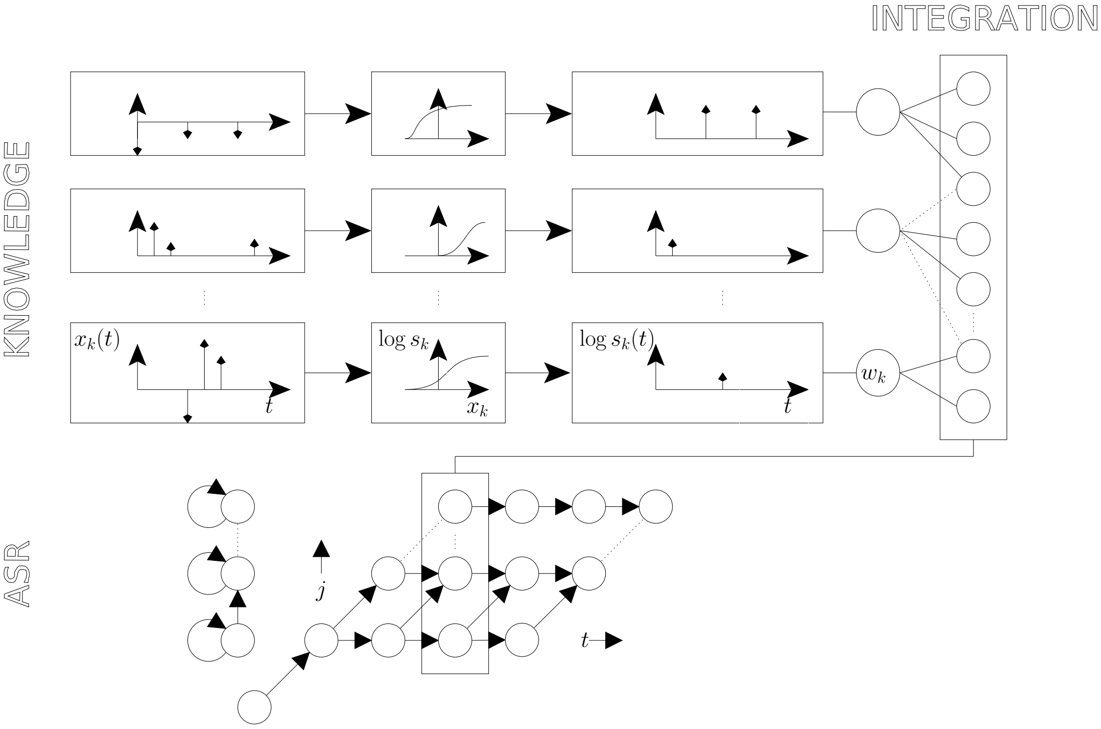


Figure 8.1.: Block diagram of the proposed integration framework, displaying k heterogeneous and asynchronous detection functions $x_k(\tau_k)$ which are mapped by the logistic mapping function onto log-likelihood scores $\log s_k(t)$, before they are combined with the standard emission probabilities of the baseline ASR system, using weights w_k.

the fact that each detection function is of arbitrary range. Since landmarks $t \in \tau_k$ appear asynchronously, this has to be done individually for each of the k detection functions. The second part weights each score $\log s_k(t)$ with a weight w_k that has been obtained by discriminatively training all k knowledge sources with the emission probabilities of the acoustic model to guarantee that detection functions provide complementary knowledge for the speech recognizer.

Mapping knowledge sources onto log-likelihood scores Each k-th input source $x_k(\tau_k)$ is mapped onto a log-likelihood $\log s_k(t)$ in order to obtain a score that is proportional to the probability that $q_{\mathcal{P}}(t) \in \mathcal{S}_k$. This score is only active, with $\log s_k(t) \geq 0$, if the phonetic class is supposed to be present, i.e., there is no negative score predicting the absence of the phonetic class under consideration. While the relation between $x_k(\tau_k)$ and $\log s_k(t)$ is non-linear, one can see that it might be reasonable to assume that mapping $x_k(\tau_k)$ onto $\log s_k(t)$ should follow a simple logistic function, that requires only few parameters to be estimated for each k-th phonetic class. First, $\log s_k(t) = 0$ for low values of $x_k(\tau_k)$, assuming a positive correlation between $x_k(\tau_k)$ and $\mathcal{S}_k$. This is straightforward, since low values $x_k(\tau_k)$ correspond to noise in the detection function and do not indicate the presence of $\mathcal{S}_k$. Second, as soon as the likelihood of $q_{\mathcal{P}}(t) \in \mathcal{S}_k$ exceeds the likelihood of $q_{\mathcal{P}}(t) \notin \mathcal{S}_k$, $\log s_k(t)$ should rise according to the growing confidence that $q_{\mathcal{P}}(t) \in \mathcal{S}_k$. Additionally, $\log s_k = 0$ for all the frames $t \notin \tau_k$, since there is no knowledge about the phones $\mathcal{S}_k$ at $t \notin \tau_k$. This effectively means that there will be no modification of the decoding if a knowledge source k is not present at frame t. Thus, $\log s_k(t)$ is very similar to $\Lambda_k(t)$ in the framework presented in Chapter 4, except that $\log s_k(t)$ includes information about the strength of the landmark. The following algorithm uses a sigmoid function as logistic mapping function and the parameters of the sigmoid are estimated for each class k by optimizing cross-entropy as the error function.

Discriminative training To modify the Viterbi decoding using the obtained scores $\log s_k(t)$ at each frame t, the k knowledge sources are combined with the scores of the acoustic model $\log s_{asr}(j,t)$ by weighted linear combination to obtain a modified acoustic likelihood $\log s(j,t)$ according to

$$\log s(j,t) = w_{asr} \log s_{asr}(j,t) + \sum_k w_k \log s_k(j,t) \tag{8.1}$$

with $\log s_k(j,t)$ being derived from $\log s_k(t)$ by

$$\log s_k(j,t) = \begin{cases} \log s_k(t) & \text{if } j \in \mathcal{J}_k \\ 0 & \text{else} \end{cases} \tag{8.2}$$

$\mathcal{J}_k$ is the state-space associated with $\mathcal{S}_k$, as defined in Section 5.1. The obtained landmark indication function $\log s_k(t)$ has to be scaled by an additional weight w_k, since the parameters for mapping $x_k(\tau_k)$ to $\log s_k(t)$ are obtained on different training instances $t \in \tau_k$ and it is not given that $\log s_k(t)$ corresponds to complementary knowledge to the emission probabilities of the acoustic model, i.e., $\log s_k(t)$ might not be able to correct errors during the search for the best word hypothesis. To obtain suitable weights w_k, the k landmark functions $\log s_k(t)$ are discriminatively trained with the emission probabilities $\log s_{asr}(j,t)$. In the following experiments, the acoustic models are unchanged, with setting $w_{asr} = 1$. Since k weights have to be optimized, optimization has to be carried out using gradient-based methods that approximate the WER by a suitable objective function. The following algorithm will use the frame-based maximum mutual information (MMI) criterion for discriminative training.

8.1.1. Objective functions

To map $x_k(\tau_k)$ onto $\log s_k(t)$ and linearly combine $\log s_k(t)$ and the acoustic emission probabilities, the parameters for the logistic mapping function and discriminative training have to be estimated according to two objective functions, which are discussed in this section.

Sigmoid parameters Formally, the mapping of $x_k\left(\tau_k\right)$ to $\log s_k\left(t\right)$ by a logistic sigmoid is calculated as follows:

$$\log s_k\left(t\right) = \begin{cases} \frac{\gamma_k}{1+\exp\left(-\alpha_k \cdot x_k(t) + \beta_k\right)} & \text{if } t \in \tau_k \\ 0 & \text{else} \end{cases} \tag{8.3}$$

Since a detection function $x_k\left(\tau_k\right)$ cannot modify the decoding at $t \notin \tau_k$, $\log s_k\left(t\right) = 0$ for all frames $t \notin \tau_k$. The parameters that have to be estimated for each detection function k are α_k, β_k and γ_k. β_k puts the sigmoid in its optimal working point, which makes the mapping invariant against the individual range of $x_k\left(\tau_k\right)$. Between low and high values of $x_k\left(\tau_k\right)$, there is either a smooth transition or a sharp step function, which can be adjusted by the parameter α_k. The parameter γ_k scales the sigmoid function $\left[1 + \exp\left(-\alpha_k \cdot x_k\left(t\right) + \beta_k\right)\right]^{-1}$ and, intuitively, $\log s_k\left(t\right)$ should be as high as possible for high values of $x_k\left(\tau_k\right)$, since those values are supposed to correctly predict the presence of the corresponding phonetic class.

The sigmoid in Equation 8.3 maps noisy, unreliable values onto values very close to zero and rounding those values to a limited precision results in $\log s_k(t) = 0$. In the following, the non-zero values of $\log s_k\left(t\right)$, i.e., the obtained landmark positions, are referred to as $\tau_{(s_k)}$, formally corresponding to the support of $\log s_k\left(t\right)$.

Cross-entropy error function The optimal landmarks $\log s_k\left(t\right)$ for knowledge source k and associated optimal parameters α_k, β_k and γ_k are obtained subject to a cross-entropy error criterion. To compute the error for $\log s_k\left(t\right)$, given a single speech utterance, it is necessary to compute reference classes $y_k\left(t\right)$ with $y_k(t) = 1$ if the reference alignment $q_{\mathcal{P}}\left(t\right) \in \mathcal{S}_k$ and $y_k(t) = 0$ if not. The error criterion has to take into account that for all $y_k(t) = 1$, there is no ceiling for the corresponding $\log s_k\left(t\right)$, since of course the higher the score for $\log s_k\left(t\right)$ at t where $y_k(t) = 1$ the more this score is beneficial for the decoding. Nevertheless, increasing $\log s_k\left(t\right)$ at few correct instances should not come at the expense of introducing erroneous landmarks at $y_k(t) = 0$. Therefore, the output $\log s_k\left(t\right)$ is not used directly for evaluation, but first converted into a probability estimate $p_k\left(t\right)$. A common solution for obtaining probability estimates from prediction scores is the use of the softmax function, which is for example used to transform the output of neural networks into probability estimates. Given the score indicating the presence of class k, with $\log s_k\left(t\right)$, as well as the score predicting the absence of k, with $\log \overline{s}_k\left(t\right)$, the softmax function estimates the probability $p_k\left(t\right)$ as follows:

$$p_k\left(t\right) = \frac{\exp\left(\log s_k\left(t\right)\right)}{\exp\left(\log s_k\left(t\right)\right) + \exp\left(\log \overline{s}_k\left(t\right)\right)} \tag{8.4}$$

$\log \overline{s}_k\left(t\right)$ is not predicted directly, since only the presence and not the absence of a phonetic class is included in the mapped log-likelihood $\log s_k\left(t\right)$. Thus, $\log \overline{s}_k\left(t\right)$ is indirectly given as 0 and can be set to

$$\log \overline{s}_k(t) = 0,\ \forall t.$$

Given $p_k(t)$, the goodness of fit between landmark $\log s_k(t)$ and true class $y_k(t)$ is calculated using the cross-entropy according to

$$CE_k(t) = y_k(t)\frac{\log p_k(t)}{N_{k,1}} + (1 - y_k(t))\,\frac{\log(1 - p_k(t))}{N_{k,0}} \tag{8.5}$$

Since some knowledge sources might have a skewed distribution, the regular cross-entropy is normalized by $N_{k,1}$ counting positive instances for which $y_k(\tau_k) = 1$ and respectively $N_{k,0}$ for which $y_k(\tau_k) = 0$. Equation 8.5 can be maximized by attributing the maximum possible probability $p_k(t) = 1$ for $y_k(t) = 1$ and minimizing the probability $p_k(t)$ for $y_k(t) = 0$. With $\log \overline{s}_k(t) = 0$, the minimum probability that can be obtained for $y_k(t) = 0$ is $p_k(t) = 0.5$. It is important to note that Equation 8.4 transforms $\log s_k(t)$ non-linearly, since $p_k(t)$ will saturate towards higher values of $\log s_k(t)$, which leads to greater sensitivity towards avoiding errors than increasing $\log s_k(t)$ at few correct instances. The optimization problem for each phonetic class k consists then in finding the parameters α_k, β_k and γ_k given $x_k(\tau_k)$ and $y_k(t)$, that maximize the accumulated cross-entropy over all frames τ_k:

$$F_{CE,k}(\alpha_k, \beta_k, \gamma_k; x_k(\tau_k), y_k(t)) = \sum_{t \in \tau_k} CE_k(t) \tag{8.6}$$

The optimization is constrained to $\alpha_k \geq 0$ and $\gamma_k \geq 0$ and Equation 8.6 can thus be rewritten as the final minimization problem:

$$\begin{array}{ll} \text{minimize} & -F_{CE,k}(\alpha_k, \beta_k, \gamma_k; x_k(\tau_k), y_k(t)) \\ \text{subject to} & \qquad \alpha_k \geq 0\ ,\ \gamma_k \geq 0 \end{array} \tag{8.7}$$

Obviously, the optimization has to be carried out over all training utterances, which is omitted in the above annotation.

8.1.2. Discriminative training

Objective functions for discriminative training penalize competing hypotheses while rewarding the reference hypothesis of training utterances. To estimate the optimal weights w_k for Equation 8.1, this thesis uses the frame-based MMI criterion which takes the following form:

$$MMI(t,n) = \log s(q(t),t)) - \log \sum_n \exp\left(\log s\left(\hat{q}_n(t),t\right)\right), \tag{8.8}$$

with $\log s(j,t))$ corresponding to the modified acoustic score according to Equation 8.1. $q(t)$ is the state sequence obtained by force aligning the reference utterance and the ensemble of all remaining hypotheses is approximated by the state sequences of the n-best list, with $\hat{q}_n(t)$, obtained by the n-best output of the speech recognizer, reordered according to acoustic likelihood. While one can increase the number of n-best hypotheses up to nearly $n = 100$ or more (see [Met05]) this thesis uses only one with $n = 1$, as it has been done for landmark-based models inside the SUMMIT speech recognizer ([MH04]). This turns the MMI criterion into corrective training [SMMN01] and the optimization problem consists then in finding the weights w_k that maximize $MMI(t,1)$ over all frames in $\tau = \bigcup_k \tau_{s_k}$:

$$F_{MMI}(w_k; q(t), \hat{q}_n(t), \{\log s_k(\tau)\}_k) = \sum_{t\in\tau} MMI(t,1) \tag{8.9}$$

The minimization problem subject to $w_k \geq 0$ thus becomes:

$$\begin{array}{ll} \text{minimize} & -F_{MMI}(w_k; q(t), \hat{q}_n(t), \{\log s_k(\tau)\}_k) \\ \text{subject to} & \qquad\qquad w_k \geq 0 \end{array} \tag{8.10}$$

8.1.3. Local optimization

With minimization of the cross-entropy according to Equation 8.7 and the discriminative training of the weights according to Equation 8.10, there are two optimization problems that have to be solved in order to obtain the corresponding parameters α_k, β_k, γ_k and w_k.

L-BFGS-B optimization Both functions are optimized using Hessian-based methods, since this allows to calculate updates individually for each parameter, leading to fast convergence. Therefore, optimization of Equation 8.7 and Equation 8.10 is performed by using the limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method. BFGS relies on a quasi-Netwon method where the Hessian matrix is estimated by a low-rank approaximation and limited-memory BFGS extends regular BFGS by approximating the Hessian matrix by few vectors that guarantee linear memory requirements. To handle constrained minimization, the following minimization uses the L-BFGS-B method, which accounts for box-constrained minimization problems.

Initialization and normalization An intuitive way of initializing w_k in Equation 8.10, is to search a uniform weight $w^{(0)}$ which scales all k sources equally, as it has been done for R_{max} in Equation 6.3. This is done by using a simple one-dimensional grid search to obtain the optimal uniform weight $w_k = w^{(0)}$ for all k by searching the weight $w^{(0)}$ that corresponds to the minimum of the objective function of the discriminative training problem.

The initial values for α_k, β_k and γ_k can be obtained from the distribution of the raw knowledge source $x_k(\tau_k)$. First, the initial parameter $\beta_{k,0}$, can be derived from the mean value $\mu_k = mean(x_k(\tau_k))$, or a related measure like the median, to put the sigmoid in its working point. While this value will not correspond to the optimal working point, it will be close enough to an optimal value to allow a quick convergence. The steepness of the slope, determined by α_k, should be proportional to the variance $\sigma_k^2 = var(x_k(\tau_k))$ to provide a smooth initial transition function. If the reciprocal of the standard deviation $\frac{1}{\sigma_k}$ and the mean μ_k are used together to initialize $\alpha_{k,0}$ and $\beta_{k,0}$, one can obtain a mean and variance normalization of the input. When the initial parameters $\beta_{k,0}$ and $\alpha_{k,0}$ result in zero-mean and unit-variance of $x_k(\tau_k)$, the lowest values of $x_k(\tau_k)$ will be mapped to values close to 0 and $\log s_k(t)$ is continuously increasing until $\log s_k(t) = 1$ for the highest values, that are supposed to correctly indicate frames with $y_k(t) = 1$. Choosing an appropriate initial scaling factor γ_k is less intuitive, but since $\log s_k(t)$ is used inside a softmax function which saturates towards high values of $\log s_k(t)$, the influence of γ_k onto the overall cost function diminishes with increasing γ_k. Therefore, choosing $\gamma_{k,0} = 1$ gives enough freedom to adapt the scaling to its optimal point. Overall, it can be seen that using a priori normalization of mean and variance of $x_k(\tau_k)$ allows uniform standard initialization for all knowledge sources by $\beta_{k,0} = 0$, $\alpha_{k,0} = 1$ and $\gamma_{k,0} = 1$.

By using normalization in connection with the proposed initialization of α_k, β_k and γ_k, the optimal parameters are invariant to multiplicative and additive scaling of $x_k(\tau_k)$, which makes the mapping of $x_k(\tau_k)$ onto $\log s_k(t)$ independent of the range of $x_k(\tau_k)$.

8.2. Experiments

Experimental setup Compared to all previous experiments, the dataset used for the experiments in this chapter was slightly modified, since all utterances containing out-of-vocabulary words were discarded from the development set, as well as from the test set. This allowed to align the reference transcriptions and n-best hypotheses obtained by the first pass of the speech recognizer to obtain $q(t)$ and $\hat{q}_n(t)$ for optimization and discriminative training. In the test case, the absence of out-of-vocabulary words does prevent detrimental errors introduced into the decoding by the language model. Optimization and discriminative training of emission probabilities and the k knowledge sources was conducted on the ESTER 2 development set, while speech recognition was conducted on the utterances of the test set, excluding all narrowband utterances.

While the previous two chapters focused on the development of landmark detection frameworks, this chapter shifts the focus towards developing a new integration scheme for external knowledge sources. Thus, the priority is not to improve detection approaches, but to study the relation between knowledge sources of different quality and the corresponding recognition performance. To be able to draw general conclusions, the following experiments make use of different degrees of oracle knowledge, i.e., the knowledge about reference alignments is used to bias the output scores of k detection functions $x_k(\tau_k)$ towards the correct solution, ranging from no use of oracle knowledge at all to using oracle knowledge to create very precise detection functions $x_k(\tau_k)$.

Oracle knowledge The phonetic classes used in the following experiments are the $k = 6$ BPCs including non-speech, as they have been used in previous experiments. The k knowledge sources $x_k(\tau_k)$ are derived from monophone HMMs according to the following steps:

1. The detection functions are based on the GMMs that model the states of the monophone models. Given a frame t of a speech utterance, a prediction score $\log x''_p(t)$ for each phone $p \in \mathcal{P}$ is approximated by taking the maximum value

c	##	vow	nas	plo	fri	app
0	1.1	0.0	0.0	0.0	0.0	0.0
2	2.87	0.0	0.0	0.0	0.0	0.64
3	3.52	0.61	0.02	0.7	0.0	1.4
4	4.15	1.35	0.51	1.21	0.22	1.91

Table 8.1.: Weights $w_k^{(c)}$ obtained after discriminative training of $\log s_k(t)$ with the acoustic model. ## represents the non-speech class.

among all $i = 3$ HMM state-based emission probabilities $\log x'''_{p,i}$:

$$\log x''_p(t) = \max_i \log x'''_{p,i}$$

2. The obtained prediction scores are converted into $k = 6$ scores for the broad phonetic classes, by approximating the score of each phonetic class with the maximum score of the corresponding phones $p \in \mathcal{S}_k$:

$$\log x'_k(t) = \max_{p \in \mathcal{S}_k} \log x''_p(t)$$

3. The continuous knowledge source $x_k(t)$ is finally obtained by normalizing all $k = 6$ classes

$$x_k(t) = \log x'_k(t) - \log \sum_k \exp\left(\log x'_k(t)\right) \tag{8.11}$$

4. Oracle knowledge is introduced into the k knowledge sources $x_k(t)$ by adding a bias c to the correct class at each frame where $y_k(t) = 1$, i.e.,

$$x_k^{(c)}(t) = x_k(t) + c \cdot y_k(t)$$

so that $x_k^{(0)}(t)$ corresponds to the unbiased continuous detection function $x_k(t)$.

5. The continuous detection functions $x_k^{(c)}(t)$ are transformed into asynchronous and sporadic sources by smoothing each function with a moving average filter and extracting the local maxima for each class k to obtain the final asynchronous landmark sources $x_k^{(c)}(\tau_k)$.

The following experiments make use of bias $c = 0$, $c = 2$, $c = 3$ and $c = 4$.

Obtained weights While Equation 8.7 converged in 10 to 15 iterations, Equation 8.10 actually converged after 3 to 4 iterations. Since each function $\log s_k(t)$ is very sparse, few landmark instances actually overlap with frames t for which $t \in \tau_{(s_a)}$ and $t \in \tau_{(s_b)}$ for two different knowledge sources a and b. Therefore, discriminative training could have been theoretically be divided into separate optimization problems for each weight w_k.

knowledge	bias	WER [dev]	WER [test]
baseline	-	28.0	31.8
BPC	c=0	28.0	31.8
BPC	c=2	27.7	31.6
BPC	c=3	27.4	31.3
BPC	c=4	26.8	31.0
vow-nas-pl	c=3	27.5	31.7
vow-nas-pl	c=4	27.3	31.5

Table 8.2.: Speech recognition results using different kinds of detection functions $x_k^{(c)}(\tau_k)$.

The weights obtained after discriminative training for all biases ranging from $c = 0$ to $c = 4$ are displayed in Table 8.1. The lower the weight w_k the less $\log s_k(t)$ actually contributed to increasing the discriminative training criterion, until the point where sources k have been completely discarded with $w_k = 0$ to avoid that $\log s_k(t)$ enhances the wrong path. This shows that discriminative training can effectively block the usage of knowledge sources that do not provide complementary knowledge to the decoding. Indeed, it can be seen that even for a high bias with $c = 2$, which corresponds to considerable oracle knowledge, only few phonetic classes provide complementary knowledge according to the MMI criterion. Figure 8.2 displays a real-world example for detection functions $x_k(\tau_k)$ and obtained log-likelihood streams $\log s_k(t)$.

8.2.1. Speech recognition experiments

Table 8.2 displays the speech recognition results obtained on the modified ESTER 2 task for the k knowledge sources using $c = 0, 2, 3, 4$ along with the baseline system. Since $w_k^{(0)}$ obtained the weight $w_k = 0$ for all phonetic classes except non-speech, the resulting WER is identical to the baseline. Despite the fact that all weights except non-speech and approximants are close to zero for all c below $c = 4$, increasing the bias improved the WER 0.2% ($c = 2$), 0.5% ($c = 3$) and 0.8% ($c = 4$). There is a considerable difference between the improvement obtained on the development set and the improvement on the test set which shows that the weights overfitted towards minimizing the error on the development set. While this is a general problem in discriminative training, the MMI objective function used in this chapter might enforce this by not using more than $n = 1$ competing hypothesis and not smoothing the error function.

When discriminative training is carried out using only the $k = 3$ phonetic classes vowels, nasals and plosives, the WER considerably increases compared to $k = 6$ phonetic classes and provides only a small improvement. While this shows the ability of the proposed approach to only use selected phonetic classes, that do not have to cover the whole phone inventory, one can see that using only few phonetic classes might not be able to correct many errors of the baseline system, since only truly complementary information propagates into the decoding and classes like plosives, nasals and fricatives

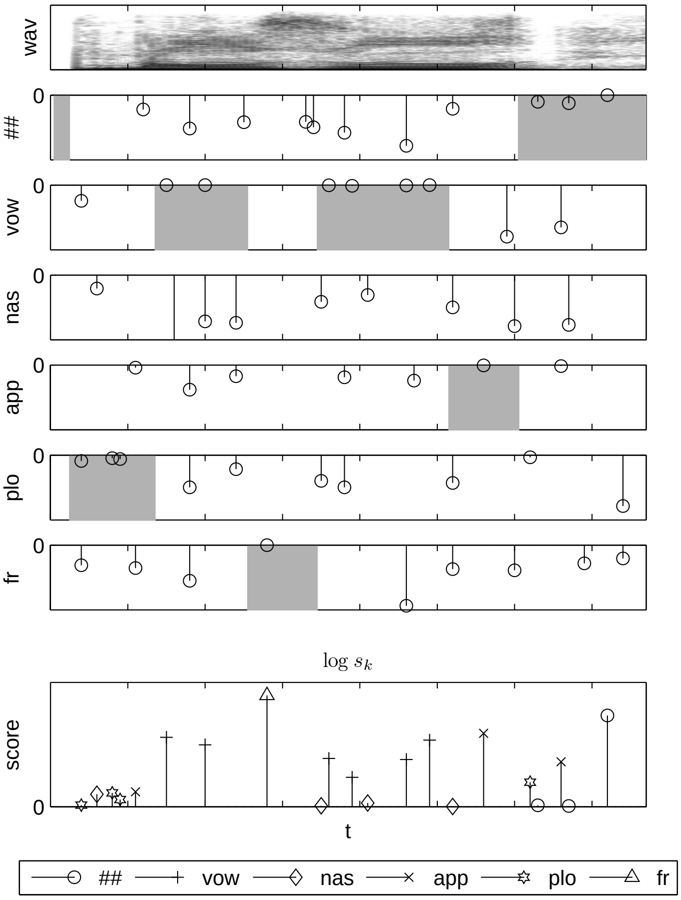


Figure 8.2.: Spectrogram of the French word «Bonjour», uttered at the beginning of a broadcast show, followed by six detection functions $x_k^{(2)}(t)$ including non-speech (##) and the obtained log likelihoods $\log s_k(t)$ at the bottom. All $x_k(t)$ are normalized, so that 0 represents the maximum value. The correct sequence of BPCs is marked in grey.

c		τ	##	vow	nas	plo	fri	app
0	AUC	τ_k	0.84	0.90	0.95	0.93	0.96	0.83
		$\tau_k^{(\neq)}$	0.41	0.43	0.37	0.35	0.34	0.46
	MI	τ_{s_k}	0.9	0.6	2.1	1.7	2.5	0.8
		$\tau_{s_k}^{(\neq)}$	0	-0.1	-0.5	-0.5	-0.8	-0.1
2	AUC	τ_k	0.94	0.96	0.98	0.98	0.99	0.93
		$\tau_k^{(\neq)}$	0.67	0.63	0.59	0.61	0.53	0.70
	MI	τ_{s_k}	1.9	1.0	3.3	3.0	3.1	1.9
		$\tau_{s_k}^{(\neq)}$	0.7	0.4	0.5	0.6	-0.2	0.6
4	AUC	τ_k	0.98	0.99	0.99	0.99	0.99	0.98
		$\tau_k^{(\neq)}$	0.87	0.81	0.80	0.83	0.73	0.88
	MI	τ_{s_k}	3.0	1.7	4.6	4.2	3.6	3.2
		$\tau_{s_k}^{(\neq)}$	1.9	1.3	2.1	2.1	0.8	2.0

Table 8.3.: Evaluation of different detection functions $x_k(t)$ and landmarks $\log s_k(t)$ using evaluation measures AUC and MI.

seem to provide very few complementary information.

8.2.2. Evaluating sporadic knowledge sources

While the relation between raw knowledge source $x_k(\tau)$ and obtained speech recognition improvement has been considered as a «black box» in previous chapters, this section aims at drawing several links between the quality of detection function $x_k(\tau_k)$ and landmarks $\log s_k(t)$ and the obtained WER. $x_k(\tau_k)$ and $\log s_k(t)$ are quantified by two different error measurements:

- The first criterion is the area under the curve (AUC) calculated on $x_k(\tau)$. AUC is a performance measurement derived from the ROC curve (receiver operator characteristic) and equal to the probability that a classifier will rank a randomly selected true BPC higher than a randomly selected false BPC. The AUC evaluates the quality of the raw knowledge source, with values above 0.9 indicating good detection performance and everything below 0.6 approaches random performance. $AUC = 1$ corresponds to perfect detection. The AUC is only evaluated on the detected frames, thus does not take missed subword units into account.

- The second criterion is a misclassification cost (MI) calculated on $\log s_k(t)$ which is related to mutual information, by calculating the average score added at each frame $t \in \tau$, weighting every correct frame with 1 and every incorrect frame by -1. This results in a negative value if a knowledge source introduces more errors into the decoding than it enhances the correct path.

$$MI(k,\tau) = \frac{1}{|\tau|} \sum_{t \in \tau} (2y_k(t) - 1) \log s_k(t)$$

$|\tau|$ corresponds to the number of frames τ over which $MI(k, \tau)$ is calculated, which is used for normalization. MI shows whether the log-likelihood scores $\log s_k(t)$ do, on average, increase the score of the correct path. Ideally, $\log s_k(t) = 0$ at frames $y_k(t) = 0$ and as high as possible for $y_k(t) = 1$. Clearly, the higher MI, the better, but values lower than 0 means that $\log s_k(t)$ rather enhances incorrect frames than the correct path.

The two criteria are calculated on two different sets of frames. First, they are calculated on all frames τ_k for $x_k(\tau_k)$ and AUC , respectively τ_{s_k} for $\log s_k(t)$ and MI. Second, they are calculated on those frames $\tau^{(\neq)}$ for which the BPC k in the reference alignment $q_{BPC}(t)$ is different from the most likely word hypothesis $\hat{q}_{BPC}(t)$ with $q_{BPC}(t) \neq \hat{q}_{BPC}(t)$.[1] For $x_k(\tau_k)$, this results in frames $\tau_k^{(\neq)} = \tau_k \cap \tau^{(\neq)}$ and for $\log s_k(t)$ this results in $\tau_{s_k}^{(\neq)} = \tau_{s_k} \cap \tau^{(\neq)}$. The values of AUC and MI calculated on different frames for different knowledge sources on the development set are displayed in Table 8.3.

By comparing the AUC for different detection functions on all frames τ_k, one can see that all detection functions perform well, even for $c = 0$, and they approach $AUC = 1$ as soon as a small bias with $c = 2$ is added. Yet, looking at the AUC at frames $\tau_k^{(\neq)}$, one can see below random performance for $c = 0$ with $AUC < 0.5$, which is slowly increasing towards higher biases. Clearly, performance on $\tau_k^{(\neq)}$ is crucial, since bad performance on $\tau_k^{(\neq)}$ propagates into a negative MI at $\tau_{s_k}^{(\neq)}$and discriminative training consequently attributes those detection functions with $w_k = 0$ to prevent the discriminative training criterion from degrading.

It is especially interesting to compare the performance of individual phonetic classes. Fricatives, who were thought of to have well identifiable acoustic cues, can only improve the discriminative training criterion with a very high bias. Approximants, which are normally associated with acoustic cues that are difficult to detect, obtain the second highest weights because of their high MI, even for $c = 2$. While the initial idea of landmark-driven ASR was to model phonetic classes which have well studied cues, like fricatives and plosives, it seems that these identifiable cues are already captured by the acoustic models and the decoding framework. Since even for a high bias like $c = 4$ fricatives do have the lowest AUC and MI value, it seems that fricatives that are not detected by the acoustic models are generally very hard to detect or cause false alarms which are difficult to avoid.

8.3. Conclusions

The presented framework allows to integrate asynchronous and heterogeneous landmarks into the decoding, by mapping each source onto logarithmic scores. Discriminative training of all knowledge sources allows to combine the mapped knowledge scores with the emission probabilities of the ASR system, by weighted linear combination. The major conclusion that can be drawn from the theoretical experiments is that landmarks

[1] Similar to $q_{\mathcal{P}}(t)$, which maps the state-alignement $q(t)$ of the most likely word hypothesis onto the phone level, $q_{BPC}(t)$ maps the alignment onto a broad phonetic level.

are useful for the decoding, when they are able to achieve above random detection performance on frames where the baseline acoustic models align the wrong path. Furthermore, one can see big differences in the ability of the individual phonetic classes to improve the discriminative training criterion and less studied speech classes, like approximants and non-speech classes, showed to provide more improvement at a lower precision, compared to well studied classes like fricatives or plosives.

Two aspects are important for further improving the presented framework. First, discriminative training should either be extended towards using a longer n-best list for modeling the competing hypotheses or it should be switched to lattice-based discriminative training. The second aspect concerns the context dependency of log-likelihoods $\log s_k(t)$. The current approach does not take preceding and following landmarks of other phonetic classes for scaling $\log s_k(t)$ into account. Yet, there are some clear dependencies between landmarks. For example, if a landmark corresponding to a plosive is immediately followed by a landmark indicating a fricative, and the score $\log s_k(t)$ of the fricative is considerably higher then the score for the plosive, it can be assumed that the score attributed to the plosive is the result of an acoustic confusion, rather than a true plosive. A simple improvement would be to fold landmarks at frame t for phonetic class k with a small window function, and weight landmarks of phonetic classes other than k inside that window dependent on the likelihood of $\log s_k(t)$.

Evidently, it is important to provide perspectives on the type of landmark detection front-ends that could be integrated using the proposed integration framework, which will be discussed in the concluding chapter.

9. Conclusions and future work

This chapter concludes this thesis by summarizing the contributions and results of all individual chapters and giving an outlook on future work.

9.1. Summary and conclusions

The general conclusion of this thesis is divided into two parts, with the first part discussing Chapter 4 until Chapter 7, which all relied on the same general landmark detection framework presented in Chapter 4. The second part of the conclusion discusses the second landmark detection framework presented in Chapter 8.

ASR driven by binary landmarks Chapter 4 presented two different landmark detection frameworks. One relying on a single front-end for each phonetic class k, the other using one shared front-end for all k classes. Yet, this thesis did not attempt to design individual detection front-ends for single classes since landmark-based pruning, which would have allowed to combine statistical ASR with an arbitrary amount of individual detection front-ends, did turn out to be too sensitive towards detection errors.

By switching from pruning to a weighted combination of binary landmarks and acoustic emission probabilities, only landmark detectors based on a shared front-end have been developed in this thesis, which all relied on training homogeneous multi-class classifiers to obtain synchronous landmarks. Two shared front-ends have been developed in this thesis and the sporadic detection functions obtained from these front-ends have been converted into a stream of binary landmarks $\Lambda_k(t)$, before integrating them into the decoding. Thus, it was never attempted to convert phonetic knowledge about individual phonetic classes and their corresponding acoustic cues into technical counterparts, but phonetic knowledge was used to design some components inside general landmark detection approaches.

In Chapter 6, these phonetically motivated components correspond to the use of a simple articulatory model for segmentation and acoustic parameters instead of MFCCs, to extract segment-based acoustic observations. Using decision trees as classification method was equally inspired by acoustic-phonetic approaches to ASR. Evaluating the landmarks obtained from the proposed system on a classification task indicated that the landmark detection framework was indeed able to avoid the modeling of acoustically ambiguous parts in the speech signal and detect phonetic classes with more precision than landmarks obtained from a simple frame-based classifier. Yet, speech recognition running modified Viterbi decoding with frame-based predictions and landmarks obtained from the proposed framework achieved similar gains in WER compared to the baseline. Chapter 7 reduces the use of phonetically motivated models but keeps a segment-based approach to landmark detection by training acoustic observations for time-variable subword units and uses the corresponding classifier to obtain a classical detection function for locating landmarks, which did improve speech recognition with regards to the baseline, but not with regards to the improvement obtained by both methods in Chapter 6. Summarizing the results from Chapters 6 and 7 results in two major points. First, while using speech segments over frames helps to increase the accuracy of acoustic observations, this segmentation does not have to be articulatory motivated, as proposed in Chapter 6. Second, frame-based phonetic predictions can compensate the lack in accuracy compared to landmarks obtained from segment-based classifiers simply by being present at every frame, instead of selected time instances, which leads to similar scores for the active hypotheses during decoding.

One reason for the rather poor performance of landmark models presented in this thesis is to be found in the use of a single front-end that always trains a multi-class classifier based on homogeneous acoustic observations. This modeling is very similar to the frame-based emission probabilities and improvements with regard to the baseline are probably rather due to using discriminative classifiers and extended acoustic observations than true phonetic knowledge, which motivated the development of the framework proposed in Chapter 8.

Integrating heterogeneous and asynchronous landmarks into ASR The framework presented in Chapter 8 attempted to integrate an arbitrary number of detection functions according to Chapter 4 into standard ASR. To account for asynchronous and heterogeneous landmarks, the k detection functions are mapped onto a stream of log-likelihood scores $\log s_k(t)$, each score representing the confidence that the landmark indicates the correct path at frame t. To integrate only complementary knowledge sources into ASR, the obtained likelihood scores are discriminatively trained by optimizing frame-based MMI. The experimental evaluation using knowledge sources of different qualities allows to draw several conclusions about landmark-driven ASR. First, phonetic landmarks might achieve high accuracy when evaluated on all frames, but only if the discriminative training criterion can be improved, the landmarks propagate into an improved word hypothesis. Some phonetic classes, like fricatives, seem to achieve low error rates on all frames, but are very difficult to detect on frames that are confused by the baseline system. Despite the use of an oracle to bias landmark detection, the improvements were relatively moderate since two points of the presented landmark detection and integration approach do additionally prevent further improvement with regard to the baseline. The first point is the use of landmarks in the first pass only, which does not necessarily guarantee that slight improvements in accuracy propagate into the final word hypothesis obtained after two passes. The second point is the limitation of using only broad phonetic classes, which can only improve the decoding if there is a confusion between different broad phonetic classes.

9.2. Future work

Several important aspects of landmark-driven ASR could not be explored in this thesis and some important future research topics are a consequence from conducted experiments.

9.2.1. Landmark detection

Since many potential landmark detection front-ends that can be integrated into Chapter 8 have already been discussed in Chapter 3, this section points out several general directions for new landmark detection algorithms that could provide complementary knowledge for statistical ASR.

Shared landmark-detection front-ends One particular step that might improve the output of shared front-ends that rely on homogeneous acoustic observations and multi-class or multi-label phonetic representations is the re-training of obtained landmark sequences by statistical models, for example Poisson process models (see [JN09a]). The sequential information could not only account for context-dependency of landmarks, but also contain information about the temporal relation between subsequent landmarks, which is still poorly represented in most approaches.

Additionally, shared-front ends for landmark detection should be extended beyond broad phonetic classes, to include place of articulation features as well as vowel features, in order to provide more information to the baseline acoustic models.

Data-driven methods for landmark detection Data-driven methods could be an alternative in case phonetic knowledge is not able to lead to accurate models for phonetic classes. Two aspects are especially interesting for the use of data-driven methods. First, instead of using natural speech classes or distinctive features to determine fixed phonetic classes, one could exploit clustering methods to cluster similar acoustic events into speech classes, which can then be detected with increased accuracy. Second, since acoustic cues for some phonetic classes, for example approximants, might be difficult to convert to precise technical models by following phonetic studies, it might be possible to experiment with data-driven methods that detect repeating patterns at phonetic events. These patterns might reveal a spectral structure that can be captured by a general model, which can then be trained using regular statistical classifiers.

Optimization criteria for landmark training Landmarks are usually trained by optimizing the frame error rate as the major error criterion. Chapter 8 showed that it is especially important to train on frames that are not correctly aligned by the ASR system. Landmark detectors can thus directly be trained to optimize error rates or ROC curves of misaligned frames or directly discriminatively trained with the emission probabilities of the acoustic models.

9.2.2. Visual knowledge

It is worth mentioning that the framework proposed in Chapter 8 is not limited to phonetic knowledge. Visual knowledge is an interesting potential additional knowledge source, which might be compatible with landmark-based models, as it has been mentioned several times throughout this thesis. Visual knowledge corresponds to information about the lip movement of speakers and certain visual gestures (visemes) can be associated with a collection of phones, which is identical to the definition of $\mathcal{S}_k$ in this thesis. Visual gestures might favor an event-based modeling paradigm, i.e., visual gestures might be recognized as single time instances by identifying perceptual relevant visual cues, e.g. lip closures or lip-rounding, using image processing. Thus, landmark-based models for visual speech gestures might reduce the noise in the visual modality of acoustic-visual ASR systems and are straightforward to integrate into ASR using the proposed integration framework.

9.2.3. Landmarks as «islands of reliability»

In this thesis, landmarks have always been defined as time instances that provide information about the presence of phonetic classes. In related work (e.g., [Sai09, LZG12]) selected time instances correspond to portions of the speech that stand out due to their intelligibility compared to their local neighborhood, for example stressed parts of the signal. To not only exploit knowledge about phonetic classes at landmarks, but also the fact that acoustic observations at landmark frames might be more reliable to predict than regular frames, landmarks could be used to increase the weight of the regular acoustic models at landmark frames during Viterbi decoding.

A. Appendix

A.1. Examples for different levels of speech transcription

Figures A.1, A.2 and A.3 display three examples for the three different types of speech transcriptions taken from [BF92], which have been discussed in Section 1.4.

A.2. ESTER 2 recognition corpus

The dataset of the ESTER 2 campaign for the rich transcription of French radio broadcasts [GGC09] consists of radio broadcasts in the French language aired during 1998 to 2008. The overall training data consists of 150 hours manually transcribed broadcast news released from 1999 to 2003 and 45 hours are taken from the EPAC project [EBA$^+$10], adding mostly non-planned speech (e.g., interviews) to the corpus. The corpus is augmented by 100 hours of transcribed radio broadcasts from the ESTER 1 corpus [GGG$^+$06].

The development set consists of 6 hours of radio broadcast news and the test set includes 7 hours of radio broadcasts recorded in January and February 2008. Both sets contain many different types of broadcast shows: broadcast news corresponding to planned, non-accented speech in studio environments with France Inter (Inter) and Radio France International (RFI), news shows containing strong accents (Africa 1 and Radio Congo) and accented interactive shows involving spontaneous speech (TVME).

The HMM-based acoustic models used in the baseline ASR system are trained on the whole ESTER 2 training corpus. Statistical classifiers for the landmark detection algorithms in Chapters 6 and 7 are trained using either 1/4 of the training data (Chapter 6) or the whole training set (Chapter 7). The official development and test set of the ESTER 2 corpus are used in a slightly modified way in this thesis, by disregarding all broadcasts with a length over 20 minutes, which was done with the sole purpose of having approximately uniform decoding time for all broadcast files during the speech recognition experiments (see Table A.1).

A.3. Baseline speech recognizer and tools

A.3.1. Baseline speech recognition

The speech recognition system used in this thesis is a classical two pass system, with a first pass generating a word graph which is rescored with more sophisticated acoustic models in the second pass. The following section gives an overview of the individual

Broadcast	date	duration
RFI	2007/07/10	20min
Africa 1	2007/06/08	15min
Africa 1	2007/06/13	15min
Africa 1	2007/06/14	15min
Africa 1	2007/06/15	15min
Africa 1	2007/06/18	15min
Africa 1	2007/06/19	15min
Africa 1	2007/06/25	15min
Africa 1	2007/06/26	15min
Africa 1	2007/06/28	15min
Africa 1	2007/06/08	15min
Inter	2007/07/10	20min
Inter	2007/07/11	20min
Inter	2007/07/12	20min
TVME	2007/07/15	15min
TVME	2007/07/16	15min
TVME	2007/07/17	15min
TVME	2007/07/18	15min

(a) development set

Broadcast	date	duration
RFI	2008/01/18	10min
RFI	2008/01/22	10min
RFI	2008/01/22	10min
RFI	2008/01/23	10min
RFI	2008/01/24	10min
RFI	2008/01/25	10min
RFI	2008/01/28	10min
Africa 1	2008/02/04	10min
Africa 1	2008/02/05	10min
Africa 1	2008/02/06	10min
Africa 1	2008/02/07	10min
Africa 1	2008/02/08	10min
Africa 1	2008/02/09	10min
Africa 1	2008/02/10	10min
Africa 1	2008/02/11	10min
Africa 1	2008/02/12	10min
Inter	2007/12/18	20min
Inter	2007/12/20	20min
Inter	2007/12/21	20min
TVME	2007/12/19	15min
TVME	2007/12/21	15min
TVME	2008/01/07	15min
TVME	2008/01/08	15min

(b) test set

Table A.1.: Files from the official ESTER 2 development and test set used for the experiments in this thesis. All recordings longer than 20mins have been discarded from the official development and test sets.

components of the baseline system. The system makes use of Spro[1] for signal processing, AudioSeg[2] for diarization, a modified version of Sirocco[3] for decoding and HTK [YEK+02] for rescoring.

Parametrization and diarization The waveform is split into fixed-size windows with a window length of 20ms and 10ms overlap For each frame 12 MFCC coefficients plus the energy are extracted. The MFCCs are normalized using running cepstral mean subtraction, with a running mean of 300ms and are additionally augmented by the first and second order derivations.

[1] available at https://gforge.inria.fr/projects/spro
[2] available at https://gforge.inria.fr/projects/audioseg
[3] available at https://gforge.inria.fr/projects/sirocco

Speech activity and narrow/wideband detection is performed using simple GMM and HMM-based models and speaker diarization is based on bottom-up clustering of the obtained speech segments based on the Bayesian Information Criterion.

Acoustic models and passes The baseline system uses two different acoustic and language models for first and second pass. The first pass uses context-dependent word-internal phones modeled by 32 state GMMs and diagonal covariance matrices with 4,019 distinct states in connection with a trigram language model. The second pass extends the acoustic models to word-external context-dependent phones modeling more than 6,000 distinct states together with a 4-gram language model. Both models are gender-dependent but do not distinguish between narrow and wideband speech. The first pass is decoded using the algorithm proposed in [NO99]. The pruning criteria used are acoustic beam, language-model pruning and histogram pruning.

Additional to the standard acoustic models used in the recognizer, some experiments make use of an additional monophone acoustic model, consisting of context-independent 64 state GMM-based HMMs with diagonal covariance matrices, equally trained on the ESTER 2 training data.

A.3.2. Tools and acoustic features

Decision trees and feature selection (Chapter 6) was performed using the WEKA toolbox [HFH+09] and boosted decision stumps relied on the tool Bonzaiboost[4]. The acoustic features used in Chapter 6 are mostly calculated using the YAAFE-Toolbox [MEF+10] and pitch and formants are calculated using the SNACK toolkit [BES98]. Table A.2 displays a list with the initial pool of acoustic features used in Chapter 6.

A.4. List of publications

- Stefan Ziegler, Bogdan Ludusan and Guillaume Gravier, Using Broad Phonetic Classes to Guide Search in Automatic Speech Recognition, in Proceedings of Interspeech 2012, Portland, USA, September 2012.
- Bogdan Ludusan, Stefan Ziegler and Guillaume Gravier, Integrating Stress Information in Large Vocabulary Speech Recognition, in Proceedings of Interspeech 2012, Portland, USA, September 2012.
- Stefan Ziegler, Bogdan Ludusan and Guillaume Gravier, Towards a New Speech Event Detection Approach For Landmark-Based Speech Recognition, in 2012 IEEE Workshop on Spoken Language Technology, Miami, USA, December 2012.
- Stefan Ziegler and Guillaume Gravier, A Framework for Integrating Heterogeneous Sporadic Knowledge Sources into Automatic Speech Recognition, in 2013 ISCA/IEEE Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 2013.

[4] available at http://bonzaiboost.gforge.inria.fr

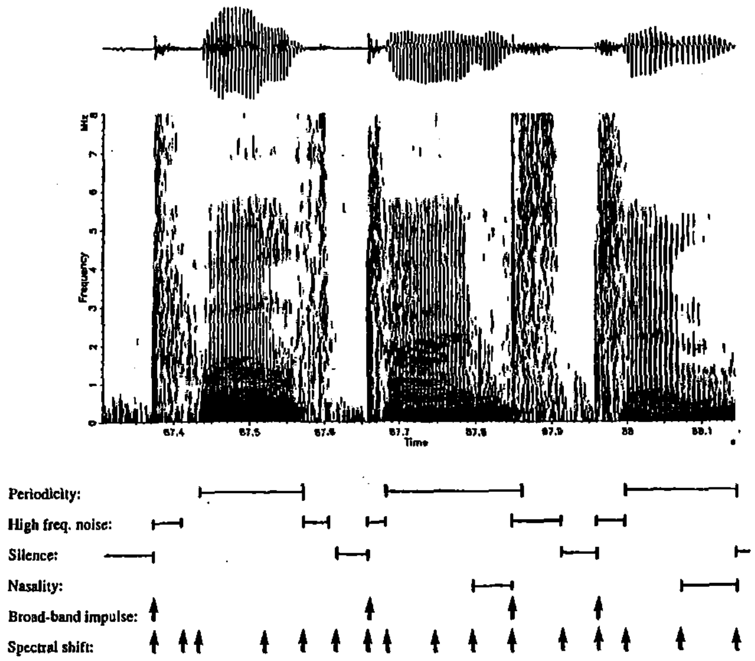


Figure A.1.: Example for a physical transcription of the utterance «twelve times ten». [BF92]

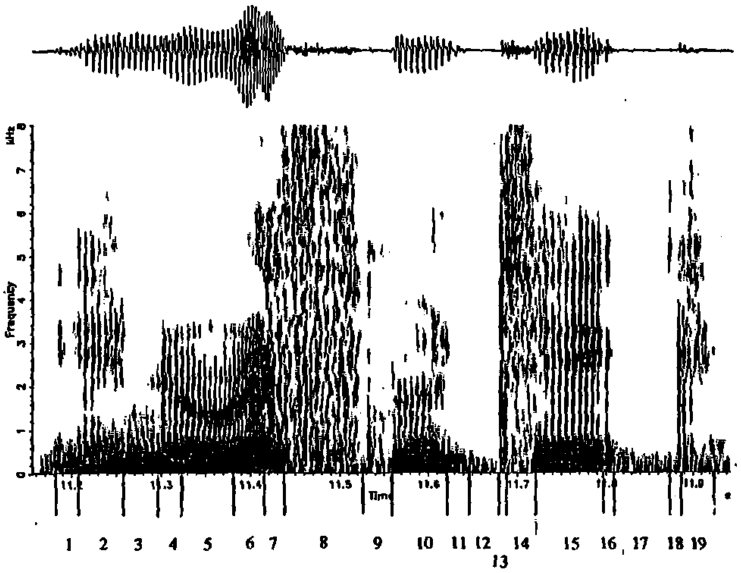


Figure A.2.: Acoustic-phonetic labels of the utterance «in arithmetic» [BF92]. While some of the labels correspond to discrete events, all labels have been attributed to segments. The annotated labels are: 1. glottal onset 2. font half-close vowel 3. nasal 4. central vowel 5. glide 6. front half-close vowel 7. voiced broad-band fricative 9. voiceless broad-band fricative 9. devoiced nasal 10. central vowel 11. voiced stop closure 12. voiceless stop closure 13. release burst 14. aspiration 15. front half-close vowel 16. glottal offset 17. stop closure 18. release burst 19. aspiration.

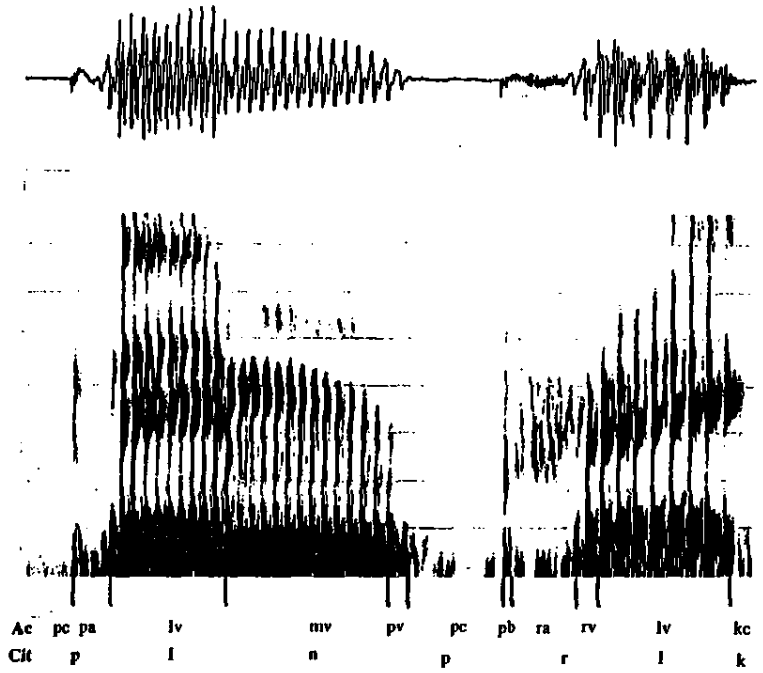


Figure A.3.: Citation-Phonemic (referred to as «Cit» in the figure) alignment of the utterance «pin prick» [BF92].

Feature	Reference
Complex Domain Onset Detection	(1)
Energy	
Envelope Shape Statistics	
Normalized Bark Bands	
Line Spectral Frequency	(2)
MFCC	
Octave band signal intensity (OBSI)	(3)
OBSIR (log-ratio of consecutive OBSI)	
Perceptual Sharpness	(4)
Perceptual Spread	
Spectral Crest Factor Per Band	
Spectral Decrease	(4)
Spectral Flatness	
Spectral Flatness Per Band	
Spectral Flux	
Spectral Rolloff	(5)
Spectral Shape Statistics	(4)
Spectral Slope	(4)
Spectral Variation	(4)
Temporal Shape Statistics	
Zero Crossing Rate	
Formant Frequency	
Formant Amplitude	
Pitch	

Table A.2.: Acoustic features used in Chapter 6 for feature selection, mostly calculated using the YAAFE toolbox [MEF$^+$10]. Acoustic features that are less common in speech processing can be looked up in the displayed references, with the indexes corresponding to (1) [DBD$^+$03], (2) [BM06], (3) [Ess05], (4) [Pee04] and (5) [SS97]. Shape statistics consist of mean, spread, skewness and kurtosis. 4 formants were calculated using the SNACK toolkit.

B. Résumé étendu

Titre : Une étude sur l'intégration de repères phonétiques dans le décodage de la parole continue à grand vocabulaire

Introduction

Alors que les systèmes de reconnaissance automatique de la parole (RAP) se sont constamment améliorés au cours des dernières décennies, les connaissances phonétiques, à savoir, les connaissances sur la production et la perception de de la parole par des humains, sont devenu presque sans importance pour les approches statistiques modernes en RAP.

La principale raison de cette ignorance des connaissances phonétiques concernant la perception humaine de la parole réside dans le fait que le phénomène n'est pas entièrement compris, les ingénieurs étant confrontés à des « boîtes noires » lorsqu'il s'agit de convertir des modèles phonétiques en système de RAP.

L'objectif de cette thèse est d'améliorer l'état de l'art en reconnaissance de la parole à grand vocabulaire en prenant en compte des connaissances phonétiques par combinaison de la modélisation acoustique traditionnelle par modèles de Markov cachés et de la modélisation par repères (*landmarks*) phonétiques. L'approche par repères phonétiques modélise la parole à certains instants, permettant ainsi de modéliser uniquement les classes phonétiques en lien avec des indices acoustiques bien connus.

Les deux questions que nous étudierons sont comment détecter les repères phonétiques et comment les intégrer dans le décodeur d'un système de RAP grand vocabulaire.

B.1. Chapitre 1 : bases de la phonétique et de la phonologie

Ce chapitre présente les bases de la phonétique et de la phonologie, domaines considérés comme essentiels pour cette thèse.

La parole est produite par filtrage par le conduit vocal d'un signal d'excitation. Les différents sons de la parole, appelés phonèmes, sont créés en faisant varier le degré de constriction créé dans le conduit vocal par les articulateurs.

Les études sur la perception de la parole humaine n'ont pas convergé vers une théorie unique et il existe plusieurs caractéristiques de la parole qui ne peuvent encore être complètement expliquée : non linéarité de la parole, segmentation en unité acoustiques, invariance des son. Les théories sur la perception de la parole humaine sont d'accord sur le fait que les êtres humains perçoivent la parole via le traitement parallèle de multiples flux d'informations hétérogènes. Les humains extraient des informations à partir de le signal acoustique en détectant des indices acoustiques, ces derniers étant des motifs spectraux qui peuvent décrits par les trois dimensions présentes dans le spectrogramme.

Le chapitre présente trois niveaux de transcription de la parole˜ : le niveau physique, le niveau acoustique - phonétique et le niveau phonémique. Les systèmes de RAP statistique repose en grande partie la transcription phonémique, divisant la parole en une séquence de phonèmes en fonction des prononciations de référence d'un dictionnaire.

B.2. Chapitre 2 : la reconnaissance de la parole

Ce chapitre introduit une architecture à l'état de l'art en reconnaissance de la parole, présentant la paramétrisation acoustique, l'architecture d'un décodeur fondé sur un modèle acoustique, un modèle de langage et un lexique.

Dans tout système de RAP, le signal de parole est tout d'abord paramétrisé sous la forme d'une séquence de vecteurs d'observations, en général un vecteur de MFCC. Le modèle de langage fournit des connaissances sur les séquences de mots sous forme de n-grammes. La RAP grand vocabulaire s'appuyant sur des unités sous-lexcicales, un lexique établit pour chaque mot les prononciations possibles.

Le modèle acoustique état de l'art est le modèle de Markov caché (MMC) qui modélise chaque unité sous-lexicale sous la forme d'un automate à états finis. Chaque état correspond à une propriété acoustique d'une unité sous-lexicale, représentée par une densité de probabilité des vecteurs d'observations.

Le décodage consiste à rechercher la séquence d'état la plus probable étant donné un énoncé. Le dispositif de reconnaissance vocale utilisé dans cette thèse utilise une approche de décodage trame-synchrone fondé sur une liste d'hypothèses actives. Cette liste est soumise à un élagage acoustique et linguistique pour en limiter la taille.

Le chapitre se termine par un résumé sur les limites de l'état de l'art en RAP.

B.3. Chapitre 3 : motivation

Ce chapitre passe en revue les avantages et inconvénients de l'état de l'art en modélisation acoustique, présente les travaux connexes ainsi que l'approche sur laquelle se fonde la thèse.

Les MMC et l'approche statistique en RAP fournissent un cadre très efficace pour l'apprentissage et le décodage dont les paramètres sont estimés à partir de transcriptions lexicales. Le décodage fait usage en parallèle du modèle acoustique et du modèle de langage en une seule passe. Cependant, l'homogénéité de l'espace d'observation, l'hypothèse d'indépendance conditionnelle et l'hypothèse markovienne sont autant de facteurs qui ne permettent pas de représenter des aspects importants du signal de parole.

Des travaux connexes ont essayé d'étendre le paradigme de modélisation par MMC, par exemple en incluant des informations articulatoires dans les MMC ou en élargissant les MMC dans le formalisme plus générique des réseaux bayésiens. Plusieurs approches ont également essayé de combiner les MMC standard avec des approches phonétiques par fusion de descripteurs ou par combinaison de systèmes.

La dernière section de ce chapitre introduit le modèle fondé sur des repères (*landmark-based model*) dans lequel la parole est représentée comme une séquence d'informations pertinentes d'un point de vue perceptuel à certains instants, informations liées au contenu phonétique.

Le chapitre conclut en présentant l'approche étudiée dans la thèse, un système fondée sur les MMC qui combine des repères phonétiques avec les modèles acoustiques.

B.4. Chapitre 4 : détection des points de repère phonétiques

Ce chapitre présente le cadre général de détection des points de repère phonétiques qui est utilisé dans ce travail.

Tout d'abord, nous donnons la définition générale de la fonction de détection des repères, telle qu'elle est utilisée tout au long de cette thèse. Les repères détectés sont corrélés avec la présence de classes phonétiques et ne sont défini qu'à certains instants précis. Ces repères sont obtenus avec des techniques de classification probabiliste, deux approches étant proposées.

La première approche repose sur une collection de fonctions de détections tandis que la seconde utilise un unique front-end pour l'ensemble des classes phonétiques. Alors que la première approche fournit un ensemble de repères hétérogènes et asynchrones, la seconde produit des repères synchrones et homogènes. Les fonctions de détections sont ensuite converties en repères phonétiques binaires.

Le chapitre se termine par la mise en place des deux stratégies d'intégration qui sont utilisées pour intégrer des repères binaires dans le décodeur du système de RAP. La première approche se fonde sur l'élagage phonétique, utilisant les repères phonétiques binaires comme un critère supplémentaire d'élagage. La seconde stratégie repose sur une combinaison de classifieurs, en combinant les modèles acoustiques avec les informations concernant les points de repère.

B.5. Chapitre 5 : utilisation des repères pour l'élagage de l'espace de recherche

Ce chapitre étend une méthode existante qui utilise des repères correspondant à des macro-classes phonétiques (voyelles, semi-voyelles, consonnes occlusives, fricatives et nasales) pour l'élagage de l'espace de recherche lors du décodage, supprimant les transitions vers des états non compatibles avec les points de repères.

Les expériences initiales avec des points de repères obtenus par un oracle ont permi une réduction considérable du taux d'erreur mot (TEM). Ces résultats ne sont cependant pas confirmés dans des conditions réalistes, i.e., lorsqu'on utilise des points de repères obtenus avec des classifieurs statistiques, les erreurs de détection des repères se propageant à l'élagage et résultant en une augmentation du TEM.

B.6. Chapitre 6 : première stratégie d'intégration

Ce chapitre présente une nouvelle stratégie d'intégration ainsi qu'une première approche pour la détection de points de repère.

L'intégration par élagage est remplacée par une combinaison de classifieurs, des repères binaires étant combinés avec les scores acoustiques pendant le décodage de Viterbi en ajoutant un facteur d'amplification fixe aux hypothèses compatibles avec les informations de point de repère présenté.

La seconde partie du chapitre présente une approche pour la détection des repères qui peut se résumer en trois étapes. Tout d'abord, le signal de parole est segmenté en segments homogènes d'un point de vue spectrale par détection des changements dans les paramètres d'un filtre auto-régressif à court terme et à long terme. En deuxième lieu, un vecteur d'observation est extrait pour chaque segment par concaténation des informations spectrales du segment et des segments voisins. La troisième étape est différente pour l'apprentissage et pour la classification. En apprentissage, le vecteur d'observation obtenu est utilisé pour former des arbres de décision servant à prédire les grandes classes phonétiques associées au segment. Dans le cas de la prédiction, le classifieur appris fournit un niveau de confiance pour chaque classe phonétique. Après détermination d'une trame représentative du segment, les scores de confiances sont transformés en masque binaire pour chaque repère.

L'évaluation expérimentale fournit des détails sur la sélection de *features* pour sélectionner les caractéristiques acoustiques utiles pour la prédiction et compare la méthode proposée avec une approche standard de classification de trames. La détection de points de repères donne de meilleurs résultats que la classification directes des trames dans une tâche de classification de phonèmes.

Les résultats en reconnaissance de la parole montrent que le système de RAP combinant les points de repères et les modèles acoustiques permet une amélioration de 23,5 % à 23,2 % du TEM en transcription de bulletins d'information. En revanche, la combinaison avec le système de classification des trames donne des résultats similaires.

B.7. Chapitre 7 : deuxieme stratégie d'intégration

Ce chapitre présente la seconde approche développée dans cette thèse pour la détection des points de repère. Le problème au cœur de ce chapitre est l'extraction d'une fonction de détection au niveau des trames acoustiques utilisant des classifieurs segmentaux. La fonction de détection obtenue est incorporée dans un algorithme classique de sélection de crête de manière à convertir une fonction de détection continue en repères binaires.

La première partie du chapitre décrit une méthode qui permet de classer des segments de parole de longueur variable en utilisant les classifieurs classiques qui nécessitent un vecteur d'observation de dimension fixe. La méthode utilise la programmation dynamique pour découper chaque segment en trois parties d'où le vecteur d'observation final est extrait. Ce vecteur d'observation est utilisée avec un algorithme qui transforme une collection de segments en une fonction de détection au niveau de la trame.

L'intégration les repères binaires ainsi obtenus au système de RAP en utilisant une combinaison comme proposée au chapitre précédent se traduit par une amélioration similaire à celle obtenue au chapitre précédent.

B.8. Chapitre 8 : l'intégration de points de repères hétérogènes et asynchrones

Le dernier chapitre présente un cadre d'intégration de repère qui permet l'intégration de points de repères hétérogènes et asynchrones. Étant donné un nombre arbitraire de fonctions de détection des repères, l'intégration est effectuée en deux étapes.

Premièrement, chaque séquence de points de repère est transformé en log-score par l'utilisation d'une sigmoïde, les séquences de scores obtenues étant combinées avec les scores des modèles acoustiques en utilisant une combinaison linéaire. Les paramètres de la fonction de transfert sigmoïde sont appris de manière optimale en utilisant un critère d'erreur fondé sur l'entropie croisée. Les poids de la combinaison linéaire sont quant à eux obtenus par apprentissage discriminant en utilisant l'information mutuelle comme critère d'optimisation.

L'évaluation expérimentale est effectuée en utilisant des fonctions de détection obtenues à partir de modèles acoustiques hors contexte, la détection étant éventuellement biaisée vers la bonne classe afin de mener des expériences avec des fonctions de détection de qualité différente.

Les résultats obtenus montrent une amélioration des performances de reconnaissance de la parole lorsque les fonctions de détection fournissent des connaissances complémentaires au niveau des trames pour lesquelles l'alignement de référence diffère de la meilleure hypothèse .

B.9. Conclusions

La thèse se termine par un résumé, des conclusions et des perspectives sur les travaux futurs. La thèse conclut que la modélisation homogène de classes phonétiques n'a pas fourni de points de repères suffisamment fiables pour améliorer par rapport à une prédiction au niveau des trames, amenant à la contribution du chapitre 8 permettant d'utiliser une modélisation hétérogène.

Les perspectives sur les travaux futurs propose de réapprendre la détection des points de repères avec des techniques d'apprentissage probabilistes, d'utiliser la fouille de données pour définir des classes phonétiques et d'utiliser le concept de repère pour la reconnaissance audiovisuelle de la parole.

Bibliography

[ACMT78] B. S. Atal, J. Chang, M. Mathews, and J. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63:1535–1555, 1978.

[AO88] R. André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 36:29–40, 1988.

[APR89] D. Autesserre, G. Pérennou, and M. Rossi. Methodology for the transcription and labeling of a speech corpus. *Journal of the International Phonetic Association*, 19(1):2–15, 1989.

[Ata83] B. Atal. Efficient coding of lpc parameters by temporal decomposition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 8, pages 81–84. IEEE, 1983.

[Aub02] X. L. Aubert. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech and Language*, 16(1):89–114, 2002.

[Bak79] J. K. Baker. Trainable grammars for speech recognition. *Journal of the Acoustical Society of America*, pages 547–550, 1979.

[BB83] M. Basseville and A. Benveniste. Sequential detection of abrupt changes in spectral characteristics of digital signals. *Information Theory*, 29(5):709–724, 1983.

[BCDM88] F. Bimbot, G. Chollet, P. Deleglise, and C. Montacie. Temporal decomposition and acoustic-phonetic decoding of speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 445–448. IEEE, 1988.

[BES98] P. Brusilovsky, J. Eklund, and E. Schwarz. Web-based education for all: a tool for development adaptive courseware. *Computer Networks and ISDN Systems*, 30(1):291–300, 1998.

[BF92] W. J. Barry and A. J. Fourcin. Levels of labelling. *Computer Speech and Language*, 6(1):1–14, 1992.

[BK99] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36(1-2):105–139, 1999.

[BM94] H. A. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer, 1994.

[BM06] T. Bäckström and C. Magi. Properties of line spectrum pair polynomials - a review. *Signal Processing*, 86(11):3286–3298, 2006.

[BQH$^+$07] I. Bromberg, Q. Qian, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. M. Siniscalchi, Y. Tsao, et al. Detection-based asr in the automatic speech attribute transcription project. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1829–1832, 2007.

[Can05] L. Canepari. *A handbook of pronunciation: English, Italian, French, German, Spanish, Portugese, Russian, Arabic, Hindi, Chinese, Japanese, Esperanto.* Lincom Europa, Munich, 2005.

[CH68] N. Chomsky and M. Halle. *The sound pattern of English.* MIT Press, 1968.

[CLA05] H. Christensen, B. Lindberg, and O. Andersen. Introducing phonetically motivated, heterogeneous information into automatic speech recognition. In *The Integration of Phonetic Knowledge in Speech Technology*, pages 67–86. Springer, 2005.

[CMH$^+$03] N. Chawla, T. Moore, L. Hall, K. Bowyer, P. Kegelmeyer, and C. Springer. Distributed learning with bagging-like performance. *Pattern Recognition Letters*, 24(1):455–471, 2003.

[Coo06] M. Cooke. A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119:1562–1573, 2006.

[DBD$^+$03] C. Duxbury, J. P. Bello, M. Davies, M. Sandler, et al. Complex domain onset detection for musical signals. In *Proceedings of the Digital Audio Effects Workshop (DAFx)*, 2003.

[DKS95] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 194–202, 1995.

[DM80] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 28(4):357–366, 1980.

[DYDA12] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[EBA$^+$10] Y. Esteve, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas. The epac corpus: Manual and automatic annotations of conversational speech in french broadcast news. pages 1686–1689, 2010.

[Eid01] E. Eide. Distinctive features for use in an automatic speech recognition system. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1613–1616, 2001.

[Ess05] S. Essid. *Classification automatique des signaux audio-fréquences: reconnaissance des instruments de musique*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2005.

[EWPJD07] C. Espy-Wilson, T. Pruthi, A. Juneja, and O. Deshmukh. Landmark-based approach to speech recognition: an alternative to HMMs. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 886–889, 2007.

[FAG08] J. G. Fiscus, J. Ajot, and J. S. Garofolo. The rich transcription 2007 meeting recognition evaluation. In *Multimodal Technologies for Perception of Humans*, pages 373–389. Springer, 2008.

[Fan71] G. Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

[Fis97] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 347–354. IEEE, 1997.

[FS95] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[FSA99] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal of the Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[FWK04] J. Frankel, M. Wester, and S. King. Articulatory feature recognition using dynamic bayesian networks. In *Proceedings of the International Conference on Spoken Language Processing*, pages CD–ROM, 2004.

[FWK07] J. Frankel, M. Wester, and S. King. Articulatory feature recognition using dynamic bayesian networks. *Computer Speech and Language*, 21(4):620–640, 2007.

[GGC09] S. Galliano, G. Gravier, and L. Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French broadcasts. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1149–1152, 2009.

[GGG$^+$06] S. Galliano, E. Geoffrois, G. Gravier, J. F. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the international conference on Language Resources and Evaluation (LREC)*, pages 315–320, 2006.

[GHP$^+$01] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington. Syllable-based large vocabulary continuous speech recognition. *Speech and Audio Processing*, 9(4):358–366, 2001.

[GKKC07] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky. Probabilistic and bottle-neck features for lvcsr of meetings. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 757–760. IEEE, 2007.

[Gla03] J. R. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152, 2003.

[GM07] G. Gravier and D. Moraru. Towards phonetically-driven hidden Markov models: can we incorporate phonetic landmarks in HMM-based ASR? In *Proceedings of the Worksop on Non Linear Speech Processing (NOLISP)*, pages 161–168, 2007.

[GMM00] D. Gibbon, I. Mertins, and R. Moore. *Handbook of multimodal and spoken dialogue systems: resources, terminology, and product evaluation.* Springer, 2000.

[GP88] J. A. Gierut and D. B. Pisoni. Speech perception. In N. J. Lass, L. V. McReynolds, J. L. Northern, and D. E. Yoder, editors, *Handbook of speech-language pathology and audiology*, pages 253–276. B.C. Decker, 1988.

[HAB$^+$90] J. Hieronymus, M. Alexander, C. Bennett, I. Cohen, D. Davies, J. Dalby, J. Laver, W. Barry, A. Fourcin, and J. Wells. Proposed speech segmentation criteria for the scribe project. *SCRIBE Project Report*, 1990.

[Hal99] M. Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

[Her90] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[HFH$^+$09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, 2009.

[HH94] W. Holmes and M. Huckvale. Why have hmms been so successful for automatic speech recognition and how might they be improved. *Speech, Hearing and Language*, 8:207–219, 1994.

[HH03] M. A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering*, 15(6):1437–1447, 2003.

[HJBB$^+$05] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang. Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 213–216, 2005.

[HKT95] W. J. Hess, K. J. Kohler, and H.-G. Tillmann. The phondat-verbmobil speech corpus. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 863–866, 1995.

[How00] A. W. Howitt. *Automatic syllable detection for vowel landmarks.* PhD thesis, Massachusetts Institute of Technology, Cambridge UK, 2000.

[Int99] International Phonetic Association. *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet.* Cambridge University Press, Cambridge, UK, 1999.

[JDM97] R. J. Jones, S. Downey, and J. S. Mason. Continuous speech recognition using syllables. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1171–1174, 1997.

[Jel97] F. Jelinek. *Statistical methods for speech recognition.* MIT Press, 1997.

[JEW08] A. Juneja and C. Espy-Wilson. A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *Journal of the Acoustical Society of America*, 123:1154–1168, 2008.

[JN08] A. Jansen and P. Niyogi. Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition. *Journal of the Acoustical Society of America*, 124:1739–1758, 2008.

[JN09a] A. Jansen and P. Niyogi. Point process models for event-based speech recognition. *Speech Communication*, 51(12):1155–1168, 2009.

[JN09b] A. Jansen and P. Niyogi. Robust keyword spotting with rapidly adapting point process models. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2767–2770, 2009.

[KCB01] J. Keshet, D. Chazan, and B.-Z. Bobrovsky. Plosive spotting with margin classifiers. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1637–1640, 2001.

[KFL$^+$07] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester. Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, 121:723–744, 2007.

[KFS02] K. Kirchhoff, G. A. Fink, and G. Sagerer. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3):303–319, 2002.

[Krs99] S. Krstulović. LPC-based inversion of the DRM articulatory model. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 125–128, 1999.

[Krs00] S. Krstulović. LPC modeling with speech production constraints. In *Proceedings of the fifth Speeech Production Seminar*, 2000.

[Lad82] P. Ladefoged. *A course in phonetics*. Harcourt Brace Jovanovich, Fort-Worth, 1982.

[LCSSK67] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological review*, 74(6):431–461, 1967.

[Lee04] C.-H. Lee. From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 109–112, 2004.

[LGB03] K. Livescu, J. R. Glass, and J. Bilmes. Hidden feature models for speech recognition using dynamic bayesian networks. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2529–2532, 2003.

[Lic52] J. C. Licklider. On the process of speech perception. *Journal of the acoustical society of America*, 24:590–594, 1952.

[Lip97] R. P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[Lis86] L. Lisker. Voicing in english: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, 29(1):3–11, 1986.

[LW11] C.-Y. Lin and H.-C. Wang. Burst onset landmark detection and its application to speech recognition. *Audio, Speech, and Language Processing*, 19(5):1253–1264, 2011.

[LZG12] B. Ludusan, S. Ziegler, and G. Gravier. Integrating stress information in large vocabulary continuous speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[ME86] J. L. McClelland and J. L. Elman. The trace model of speech perception. *Cognitive Psychology*, 18(1):1–86, 1986.

[MEF$^+$10] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. YAAFE, an easy to use and efficient audio feature extraction software. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[Met05] F. Metze. *Articulatory features for conversational speech recognition.* PhD thesis, Karlsruhe Institute of Technology, 2005.

[Met06] F. Metze. Articulatory features for 'meeting' speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2006.

[MH04] E. McDermott and T. J. Hazen. Minimum classification error training of landmark models for real-time continuous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 937–940. IEEE, 2004.

[MHP12] A.-r. Mohamed, G. Hinton, and G. Penn. Understanding how deep belief networks perform acoustic modelling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4273–4276. IEEE, 2012.

[MPR02] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88, 2002.

[MW02] F. Metze and A. Waibel. A flexible stream architecture for ASR using articulatory features. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2002.

[NJRT12] A. Norouzian, A. Jansen, R. C. Rose, and S. Thomas. Exploiting discriminative point process models for spoken term detection. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2442–2445, 2012.

[NO99] H. Ney and S. Ortmanns. Dynamic programming search for continuous speech recognition. *Signal Processing Magazine*, 16(5):64–83, 1999.

[NS97] L. Nguyen and R. M. Schwartz. Efficient 2-pass n-best decoder. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*. Citeseer, 1997.

[OM78] G. C. Oden and D. W. Massaro. Integration of featural information in speech perception. *Psychological review*, 85(3):172–191, 1978.

[Ost99] M. Ostendorf. Moving beyond the beads-on-a-string model of speech. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 79–84, 1999.

[Pee04] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 2004.

[PEW04] T. Pruthi and C. Y. Espy-Wilson. Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43(3):225–239, 2004.

[PL07] D. B. Pisoni and S. V. Levi. Representations and representational specificity in speech perception and spoken word recognition. In G. Altmann and M. G. Gaskell, editors, *The Oxford handbook of psycholinguistics*, pages 3–18. Oxford University Press, 2007.

[PNLM04] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing*, 22:23, 2004.

[Rab89] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)*, 77(2):257–286, 1989.

[RDL10] D. Ruinskiy, N. Dadush, and Y. Lavner. Spectral and textural feature-based system for automatic detection of fricatives and affricates. In *Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, pages 771–775. IEEE, 2010.

[Rep82] B. H. Repp. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92(1):81–110, 1982.

[RJ93] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. 1993.

[Sai09] T. N. Sainath. Island-driven search using broad phonetic classes. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 287–292, 2009.

[SC12] G. Saon and J.-T. Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine*, 29(6):18–33, 2012.

[SK86] K. Shirai and T. Kobayashi. Estimating articulatory motion from speech wave. *Speech Communication*, 5(2):159–170, 1986.

[SL09] S. Siniscalchi and C. Lee. A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*, 51(11):1139–1153, 2009.

[SLY11] F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 437–440, 2011.

[SMMN01] R. Schlüter, W. Macherey, B. Müller, and H. Ney. Comparison of discriminative training criteria and optimization methods for speech recognition. *Speech Communication*, 34(3):287–310, 2001.

[SMSW03] S. Stüker, F. Metze, T. Schultz, and A. Waibel. Integrating multilingual articulatory features into speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1033–1036, 2003.

[SRS02] A. SaiJayram, V. Ramasubramanian, and T. Sreenivas. Robust parameters for automatic segmentation of speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 513, 2002.

[SS97] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1331–1334. IEEE, 1997.

[SS99] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[Ste02] K. N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111:1872–1891, 2002.

[VDKM89] A. M. Van Dijk-Kappers and S. M. Marcus. Temporal decomposition of speech. *Speech Communication*, 8(2):125–135, 1989.

[Wak79] H. Wakita. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art. *Acoustics, Speech and Signal Processing*, 27(3):281–285, 1979.

[YEK$^+$02] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK book. *Cambridge university engineering department*, 2002.

[YOW94] S. J. Young, J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312. Association for Computational Linguistics, 1994.

[ZHJB04] Y. Zheng, M. Hasegawa-Johnson, and S. Borys. Stop consonant classification by dynamic formant trajectory. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2004.

[ZN09] G. Zweig and P. Nguyen. A segmental CRF approach to large vocabulary continuous speech recognition. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 152–157. IEEE, 2009.

[ZR98] G. Zweig and S. Russell. Speech recognition with dynamic bayesian networks. *Proceedings of American Association of Anatomists (AAA)*, pages 173–180, 1998.

[ZTK07] H. Zen, K. Tokuda, and T. Kitamura. Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, 21(1):153–173, 2007.