Karlsruhe Institute of Technology

Unsupervised Phoneme Segmentation of Previously Unseen Languages

Master's Thesis of

Marco Vetter

at the Institute for Anthropomatics and Robotics
of the Department of Informatics

First reviewer:	Prof. Dr. Alex Waibel
Second reviewer:	Prof. Dr. Walter Tichy
Advisors:	Dr. Sebastian Stüker
	M.Sc. Markus Müller
	Prof. Satoshi Nakamura

Duration: 12. November 2015 – 11. May 2016

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 11. Mai 2016

Abstract

In this work we investigate the automatic detection of phoneme boundaries in audio recordings of previously unseen languages. The work is motivated by the need to aid the documentation of endangered languages using natural language processing (NLP). This task requires the automatic phonemic transcription of both unseen and possibly unwritten languages, of which the initial segmentation of the audio attempted here is a first, necessary step, before moving on to the classification of the segments' phonetic contents. In order to achieve this we employ monolingual and multilingual phoneme recognizers in a cross-lingual fashion. After generating the necessary boundaries, we then proceed to measure the quality of the segmentations using precision, recall and F_1-score. We compare the scores achieved by different configurations of recognizers on both English, where we also compare results to a gold standard, as well as Basaa, a Bantu language spoken on parts of the African continent.

Zusammenfassung

In dieser Arbeit untersuchen wir die automatische Erkennung von Phonemgrenzen in Audio-Aufnahmen von bis dato nicht betrachteten Sprachen. Die Arbeit ist motiviert von der Notwendigkeit, die Dokumentation von bedrohten Sprachen mit Mitteln des Natural Language Processing (Verarbeitung natürlicher Sprache, NLP) zu unterstützen. Diese Aufgabe erfordert die automatische Transkription von Sprachen auf der Phonem-Ebene, wobei die Zielsprachen bislang nicht untersucht wurden und unter Umständen auch über keine Schriftform verfügen. Die initiale Segmentierung der Audio-Daten, mit welcher sich diese Arbeit beschäftigt, ist dafür ein notwendiger erster Schritt, bevor mit einer Klassifikation der Segment-Inhalte begonnen werden kann. Um diese Segmentierung zu erstellen, verwenden wir mono- und multilinguale Phonem-Erkenner über Sprachgrenzen hinweg. Nach der Generierung der notwendigen Segment-Grenzen fahren wir fort, die Qualität der Segmentierungen mittels Precision, Recall und F_1-Score zu messen. Wir vergleichen die Ergebnisse verschiedener Konfigurationen von Erkennern auf Englisch als Zielsprache untereinander und gegen einen Gold-Standard, sowie untereinander auf der zweiten Zielsprache Basaa, einer Bantu-Sprache welche auf Teilen des afrikanischen Kontinents gesprochen wird.

Acknowledgements

I wish to thank Prof. Dr. Alexander Waibel and Prof. Satoshi Nakamura for the opportunity to conduct parts of the research that went into this work at the Nara Institute of Science and Technology.

I also wish to thank my advisors, Dr. Sebastian Stüker and Markus Müller, for their support of my work and their patience in the face of my ceaseless questions.

Contents

1. Introduction

The field of automatic speech recognition (ASR) has its roots in the second half of the 20th century. Often considered pioneers of the field, in 1952 Davis et al. devised a circuitry at Bell Laboratories that could successfully recognize single-speaker digits using the first and second formant of the speech signal [DaBB52]. The system shared with other early approaches the requirement that the signal consist solely of one complete entity that was to be recognized, e.g. a single digit, with no other speech sounds preceding or following it. The first work that successfully utilized segmentation as part oft the recognition process was published by Sakai et al. in 1962 [SaDo62].

The advent of Linear Predictive Coding (LPC) with its simplified vocal tract model in the 1970s ultimately allowed for more advanced systems and lead to the realization of speaker-independent recognition, at least of isolated words, as presented e.g. by Rabiner et al. in 1979 [RLRW79]. Another novel concept introduced during that decade was the use of graph search utilizing search beams and finite state networks, the result of which was the "Harpy" system developed at the Advanced Research Projects Agency (ARPA) in the United States of America [Lowe86].

During the 1980s researchers turned away from template-centric approaches and towards statistical methods, which finally lead to the application of Hidden Markov Model to the task of speech recognition, e.g. by Levinson et al. in 1983 [LeRS82]. The use of such HMMs to account for the variability of speech signals continues to the current day.

The 1980s also marked a resurgence of the use of artificial neural networks (ANN) in speech recognition, driven by the increase in computational power since initial attempts in earlier decades. However, success of the approach was limited due to the temporal variability of speech signals, which was only properly addressed with the introduction of Time-Delay Neural Networks in by Waibel et al. in 1989 [WHHS$^+$89]. Alternatively, ANNs were later combined with HMMs in so-called hybrid approaches due to their capability of accounting for said temporal variability.

Early target applications for speech recognition included "voice-activated typewriters" (VAT) and command and control functionality in telecommunication. The initially

speaker-dependent VATs later transitioned into speaker-independent dictation software that worked without being "trained" with voice-samples of the intended speaker prior to actual use. The most common and visible application of ASR systems today is as an additional, usually mostly speaker-independent, input modality for personal computers and other personal electronic devices. Mobile phones, tablets and car navigation systems are all examples of everyday use devices that can greatly benefit from the option for hands-off interaction. Especially mobile devices offer speech recognition as input for both dictation and command purposes. Commonly known examples at the time of writing include Apple's Siri, Microsoft's Cortana and Android's built-in speech recognizer.

Apart from commercial use, ASR can also find application in other scientific fields. The purpose of the work presented here is to support efforts to conserve the diversity of human language, i.e. to help document small and exotic languages that are threatened by extinction due to their dwindling numbers of live speakers and/or lack of an existing writing system.

1.1 Goals

According to the *Ethnologue* [14], there are currently over 7000 languages that are still spoken in the world, many of which have not only very small, but also rapidly dwindling numbers of active speakers and are therefore threatened by extinction in the near future ([NeRo00], [Crys00]). Documenting the cultural heritage that these languages represent is often hindered by a number of factors, such as time investment, lack of data and cost of acquiring data, as well as, in some cases, the complete lack of a written representation that could be used for documentation in the first place.

The first factor, the necessary time investment of trained linguists, could be mitigated by applying machine learning tools generally used in natural language processing (NLP) to the task. However, this runs afoul of the second issue, the lack of training data and the cost associated with acquiring it. State-of-the-art NLP systems require large amounts of training data in order to produce accurate results. While these *corpora* are widely available for major languages such as English, French or Spanish, they are often non-existent for small, exotic languages, with little to no economic incentive to gather and compile the necessary data. Consequently, while speech recognizers are now comparatively easy to train for wide-spread, economically significant languages, this is not the case for, e.g., indigenous or insular languages that are only spoken by a very small local speaker group, and therefore pose no incentive to build such systems.

The BULB project [ASADA$^+$16] aims to facilitate the documentation of such small, under-represented languages, especially in the case of the absence of a writing system. The project uses a multi-step approach of first collecting, re-speaking and orally translating target language data, and then processing the generated corpus, providing phonetic and word level transcriptions, as well as inter- and cross-lingual alignments in order to support the manual documentation work performed by linguists.

In this work, we address the first step necessary for this process, the phoneme segmentation. As we assume no prior knowledge of either the target language's

phonetic characteristics, grammatical structure, dictionary or writing system, we will employ cross- and multilingual phoneme recognizers to generate these segmentations. We will focus exclusively on the *positions* of the phoneme boundaries, leaving the next step - identifying and labelling the segments - for future work.

We will then compare the performance of the various cross- and multilingual recognizers amongst each other and, where possible, against a gold standard provided by a monolingual system trained on the target language.

1.2 Overview

Chapter 2 will introduce basic concepts for the methods used in this work, from the fundamental rationale behind automatic speech recognition to acoustic and language models in general, as well as the particular types of these models used. Chapter 3 briefly describes the employed framework, and gives details about the corpora used for training and tests. We will then look at the results of our experiments alongside their interpretation and discussion in chapter 4. Finally, chapter 5 presents a Summary and an outlook on possible future work.

2. Fundamentals

2.1 Automatic Speech Recognition

Automatic speech recognition (ASR) as a subject of research aims to apply the principals and algorithms of machine learning to the task of transforming an incoming human speech signal into a machine-readable representation of the word sequence encoded in said signal. In order to achieve this, the signal must be transformed into a parametric representation that is adequate for the intended purpose, i.e. in this case, the acoustic modeling.

In a first step this process involves the digitization of the analogue acoustic waveform in order to make the data properly machine-readable. This involves discretization of the signal along the time axis (sampling) as well as the value axis (quantization). As the resulting time-domain representation is not necessarily useful for the purposes of ASR, it is usually transformed into the frequency domain. After additional preprocessing steps, such as further transformations, dimension reduction or stacking, the end result will be a succession of multidimensional so-called *feature vectors*, representing the state of the original signal over time in a manner that is conducive to the task of ASR.

The dominant approach to ASR today is a statistical one. Therefore it is the task of a decoder to find, given a specific series of input feature vectors (i.e. an utterance), the most likely sequence of words encoded in the speech signal those vectors represent. Equation 2.1 shows a mathematical formulation of the problem.

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|X) = \underset{W}{\operatorname{argmax}}\, \frac{P(X|W) \cdot P(W)}{P(X)} = \underset{W}{\operatorname{argmax}}\, P(X|W) \cdot P(W) \tag{2.1}$$

$\hat{W}$ denotes the most likely word sequence, which equates to the specific sequence for which the conditional probability $P(W|X)$ is maximized, X being the observed signal. Using *Baye's Theorem* this can be expressed as the product of the conditional

probability $P(X|W)$ of an observation given a word sequence and the prior probability $P(W)$ of the word sequence itself, divided by the non-conditional probability $P(X)$ of the observation.

Since over the course of evaluating the most likely word sequence for a particular observation the observation is constant, it is generally omitted for that purpose. This leaves the likelihood $P(X|W)$, referred to as the *acoustic model*, and the probability $P(W)$, called the *language model*. The former models the probability that a specific observation (i.e. a string of input vectors obtained via pre-processing) is made given the original word sequence it represents, using knowledge of the phonetic structure of the targeted language. The latter gives the probability of that word sequence occurring in the first place, independent from any observations made, incorporating structural information on the language.

From this it becomes obvious that models trained on speech from one language will usually have limited applicability for decoding that of another. All languages differ with regard to their phonetic and grammatical properties (as well as their dictionaries), depending on their ancestry and present degree of relatedness, which is the cause of their mutual unintelligibility. Like any code, a human speaker cannot parse a language whose rules and characteristics they are not familiar with. Likewise, a speech recognizer will be unable to reliably recognize and transcribe a speech signal from a language that its models were not trained on. However, in order to enable the application of ASR to languages that are not sufficiently sourced to train such models for them specifically, there are approaches to circumvent these limitations with multilingual models. We will briefly introduce how we train such models in section 3.4.2. Sections 2.1.1 and 2.1.2 below will describe in more detail the concepts of acoustic and language models.

2.1.1 Acoustic Models

With the acoustic wave of the speech signal as the central piece of data for recognition, it follows that the quality of the acoustic model, i.e. the model that incorporates the available knowledge about the phonetic properties of the target language, is also of central importance.

The predominant approach to modeling the acoustic properties of a language are hidden Markov models (HMM). A major advantage of HMMs is that they can be used to build models in the absence of complete information about the internal state of the process that produces our observations. This is the case in speech recognition, as we can only observe the emitted sound waves, not the state or parameters of a system that produces them.

Apart from the problem of incomplete information about the process generating the signal, the variability of the signal itself is another major complication when building acoustic models. Syntactical and semantical context, style, domain, speaker and acoustic environment all vary greatly between utterances. Having to generalize across one or more such factors often heavily and negatively impacts the performance of a system. Higher robustness, for example regarding individual speakers (i.e. speaker-independence), is achieved by selecting training data gathered from a sufficiently large pool of speakers.

A well-explored and commonly used generative approach to modeling the acoustics of a language are Hidden Markov Models, which will be described in detail in the following section.

2.1.1.1 HMM-based acoustic models

Hidden Markov models are an extension of a simpler model known as the Markov chain. Markov chains can model a system using sets of possible *states* and *transitions*, assigning probabilities to each transition from a state s_i to another state s_j for each time index t, given the previous n states (n^{th}-order Markov chain). For interpreting a speech model we assume that time indices are discrete (we can do this due to the discretization of the signal mentioned in section 2.1), and that only the immediately preceding state influences the model's behaviour at any given time. We also assume that the process of speech production is time-invariant, i.e. that the probability to transition from s_i to s_j is the same independent of the time index t. Given these assumptions we refer to the resulting model as a homogenous, time-discrete 1^{st}-order Markov chain.

In actuality, the current state of a complex system, such as speech production, is often not observable. This is accounted for by extending the Markov chain with an additional element, the so-called emission probability densities. An emission probability density b_i for state s_i will specify which of the possible signals will be emitted by the modelled system when that state is entered.

With this a hidden Markov model can be formally defined as the following quintuple:

- $S = \{s_1, ..., s_N\}$ - A set of states, with q_t denoting the state at time t
- $O = \{o_1, ..., o_M\}$ - A set of observable symbols
- $A = \{a_{ij}\}$ - A transition probability matrix, with a_{ij} the probability to transition from state s_i to s_j: $a_{ij} = P(s_t = j | s_{t-1} = i)$
- $B = \{b_i(k)\}$ - Emission probability distributions, with $b_i(k)$ the probability of emitting symbol o_k when state i is entered: $b_i(k) = P(o_k | q_t = s_i)$
- $\pi = \{\pi_i\}$ - An initial state distribution with $\pi_i = P(q_1 = s_i), 1 \leq i \leq N$

In order to robustly train the parameters of a HMM, sufficient training data is paramount. Training on a word level would therefore require exorbitant amounts of data in order to assure enough samples for each word, and would also result in large numbers of models, while still running danger of encountering unseen words after training. The phoneme set of a language on the other hand is much smaller than its dictionary, and each phoneme will occur with a relatively high frequency in most reasonably sized corpora. Separating even further, for the purpose of acoustic modeling, HMMs are usually applied on a sub-phoneme level to account for the shifting characteristics of individual phonemes over time. For a three-state sub-phoneme model for example, a phoneme p would be modeled with sub-phoneme states $p-b$, $p-m$ and $p-e$.

Different topologies can be applied to these states in order to define legal transitions between them. For example, the common three-state Bakis topology will allow

transitions within the phoneme's three-state model from any state to the succeeding state, the state after that, or, in a recursion, the current state itself. The latter property allows HMMs to account for temporal distortions, such as lengthening or shortening of individual sounds in spoken language, which is one of the reasons they are so suited to the application of acoustic modeling. Finding the optimal parameters for an HMM-based recognizer (from training data) is generally done using either Viterbi or Baum-Welch training.

It should be noted that in actual speech, co-articulation effects can lead to variations in the pronunciation of phonemes, such as devoicing or aspiration. In order to account for these effects we can use so-called *context-dependent* acoustic models, a common variant of which are *triphone* models that consider one unit to either side. E.g. for a phoneme-level model (i.e. without using sub-phonemes as described above), this would mean for the voiceless glottal fricative /h/ to be represented by several polyphones: /h/(sil,/e/) when preceded by silence and followed by /e/, /h/(sil,/i/) when preceded by silence and followed by /i/, etc. In similar fashion we can construct even more complex models like quintphones (two units of context to either side) or subtriphones (one unit of context to either side, but on a sub-phoneme level).

While these models more accurately represent the reality of speech production, they also suffer from the major downside of requiring very large amounts of training data in order to ensure that every polyphone constructed from the basic phonetic inventory is covered to a degree that will allow for robust estimation of model parameters.

2.1.2 Language Models

While acoustic models, as described in section 2.1.1, represent the phonetic properties of a language, language models attempt to do the same for its grammatical characteristics, i.e. syntax and semantics. They thereby - if trained appropriately - introduce additional information about said language, making a speech recognition system that uses both models simultaneously more powerful with regard to its accuracy.

Language models are especially necessary when faced with acoustically indistinguishable, but semantically distinct speech, as in the case of homophones. Acoustically two words like <cite> and <sight> may realize the same, but they obviously have entirely different meanings. A recognizer using purely acoustic models has no way of distinguishing these words, then, which will inevitably introduce errors.

A second reason to employ language models is to reduce the search space while decoding. Acoustic models are generally trained with very narrow context, in case of a context-dependent system, or none when building a context-independent one. It is not feasible to use these models to ensure legality of phone sequences with respect to a specific language above the word level, even though obviously not any possible sequence will be possible when taking that language's grammar into account. Language models do exactly this, thereby allowing considerable reduction of the search space - the potential, not-yet-discarded hypotheses - during decoding.

Usually language models are applied to this effect on a word level. In this work, however, in an attempt to improve the performance of a phoneme segmenter, we will apply them in the same manner, but on a phonetic level. That is, instead of using

the model to estimate probabilities of specific word sequences, we will train them to provide probabilities of specific phoneme sequences. Since the acoustic models we train will be context-independent, this will allow us to still take context into account, based on the parameters set for the language model.

In the following section we will briefly introduce the fundamental concepts of the type of model chosen for this task, the n-gram based language model.

2.1.2.1 N-gram language models

Early attempts at formalizing deterministic language models employed rule-based grammars, using finite sets of terminal and non-terminal symbols alongside rewrite rules in order to model the structural properties of a natural language. The downsides of this approach are the required initial effort of defining the sets of symbols and the rules that govern the transformations between them, as well as the lack of flexibility when encountering previously unseen syntactical structures and words. These issues are much easier addressed by stochastic language models, which can be trained automatically on available corpora, and can more easily deal with previously unseen data.

A stochastic language model attributes probabilities to word sequences, generally on a sentence or utterance basis, based on their frequencies in the training data. Equation 2.2 shows how the probability of a sequence W is decomposed into a product of probabilities of individual words w_i, conditional on their respective histories, i.e. the words preceding them.

$$\begin{aligned} P(W) &= P(w_0, w_1, w_2, ..., w_n) \\ &= P(w_0) \cdot P(w_1|w_0) \cdot P(w_2|w_1, w_0) \cdot ... \cdot P(w_n|w_{n-1}, w_{n-2}, ..., w_0) \end{aligned} \tag{2.2}$$

Without further refinement this approach will be limited to assigning probabilities to only those specific word sequences that appear in the training data. To solve this, we can introduce equivalence classes that each substitute a set of histories with a specific class. Equivalence classes can, for example, be based on the syntactic roles of phrases and words, or their semantic meaning. However for the purpose of automatically training a model on a text corpus these options are not feasible, as they require additional annotation. In addition, when transferring the principles of word-level language models to the phonetic level, syntax and semantics simply do not apply as concepts.

Therefore we will turn to another, widely used type of equivalence class, called n-grams. Under the Markov assumption we can equate the probability of a specific word w_i given it's history h_i to the probability of w_i given just its immediate predecessor:

$$P(w_i|h_i) = P(w_i|w_{i-1}, w_{i-2}, ..., w_1, w_0) \approx P(w_i|w_{i-1}) \tag{2.3}$$

The equivalence class shown in equation 2.3, taking into consideration only the immediately preceeding word, is designated a *bi-gram*. Adding additional information to the model by extending the considered history the class can be extended to

tri-grams, *4-grams* etc. While higher-order n-grams make use of a wider context, thereby incorporating syntactic and semantic knowledge over a greater distance between words, they also require a bigger amount of training data in order to reach useful estimates for their probabilities in future, unseen data. In other words, a large context width that results in a relative frequency of either 0 or 1 for each word given a specific history is not useful. An order of n-grams that has been proven to be both useful and reasonable are tri-grams, incorporating a two-word history, which we are also going to use in our experiments presented here.

Training of an n-gram language model is done via maximum likelihood estimation:

$$P(w|h) = \frac{count(h, w)}{count(h)} \tag{2.4}$$

To address the issue of unseen phrases being assigned a probability of zero, we will have to apply a smoothing technique. The simplest approach, increasing the count of *all* possible events by one, also referred to as "add-one smoothing" or "Laplace smoothing". However with increasing vocabulary sizes (and context) the number of possible phrases grows quickly, which leads Laplace smoothing to attribute too much probability mass to unseen (and rarely seen) phrases.

The method we choose for our experiments is the Witten-Bell smoothing [WiBe91], which is itself a type of recursive interpolation, or Jelinek-Mercer smoothing. Jelinek-Mercer smoothing interpolates the probabilities provided by lower-order models with those of higher-order models, using a parameter λ to indicate the respective weights:

$$P_{JM}(w_i|w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} p_{ML}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda_{w_{i-n+1}^{i-1}}) p_{JM}(w_i|w_{i-n+2}^{i-1}) \tag{2.5}$$

That is, the probability of the n^{th}-order Jelinek-Mercer-smoothed model is the sum of the n^{th}-order maximum likelihood probability estimate and the $(n\text{-}1)^{th}$ order Jelinek-Mercer-smoothed model probability. The recursion ends with the interpolation of the 1^{st} order model (the unigram) and the "0^{th}-order" uniform distribution.

For Witten-Bell smoothing we use the number of possible extensions N_{1+} of each history in order to estimate the probability that a history $w_1, ..., w_{n-1}$ is followed by a word w_n when the whole sequence $w_1, ...w_n$ has not been encountered during training.

$$N_{1+}(w_1, ..., w_{n-1}, *) = |\{w_n : c(w_1, ..., w_{n-1}, w_n) > 0\}| \tag{2.6}$$

We then calculate the λ necessary for the recursive interpolation as follows:

$$1 - \lambda_{w_1,...,w_{n-1}} = \frac{N_{1+}(w_1, ..., w_{n-1}, *)}{N_{1+}(w_1, ..., w_{n-1}, *) + \sum_{w_n} c(w_1, ..., w_n)} \tag{2.7}$$

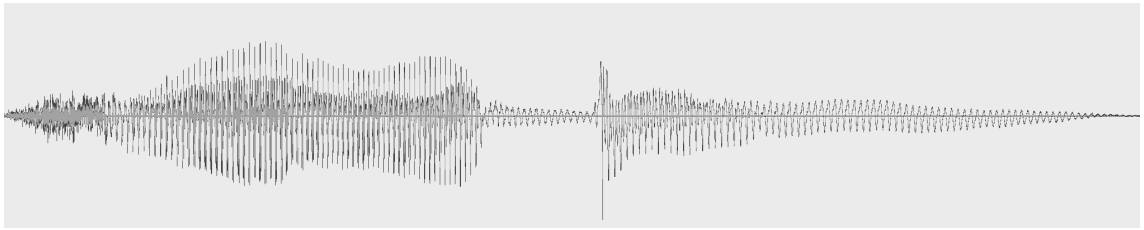

Figure 2.1: Waveform diagram for an instance of the German word ”haben”

2.2 Supervised/Unsupervised Training and Segmentation

The problem of unsupervised phoneme segmentation has been the subject of research for some time. The fundamental difficulty in achieving useful results when segmenting audio into any kind of sub-units is the continuous nature of the acoustic signal - unlike most written representations of speech, the waveform that transmits the encoded speech does not come with convenient whitespace between words, or any sort of discretization at all in the way that letters roughly segment individual words into their phonetic constituents. Figure 2.1 shows an example waveform diagram for the German word ”haben”. We can see that while some features allow for a rough division of the phoneme string, e.g. the sudden changes in amplitude before and after release of the plosive mid-word, it is not immediately apparent where exactly the border is between a voiceless glottal fricative and an open front vowel. To add to this, co-articulation effects that occur between certain neighbouring sounds can blur the segmentation lines even further, as they influence the positioning of the speech apparatus and thus lead to variations in the acoustic profiles of the produced sounds.

While the human brain is trained from infancy to segment acoustic input in order to derive the encoded information, it has proven non-trivial to impart the same ability to an artificially built recognizer (or segmenter, for that matter) running on a computer. Nevertheless, automatic speech recognition (as well as other speech technologies, such as text-to-speech) is dependent on correctly dividing the incoming speech recordings or live speech.

This segmentation is usually achieved by the acoustic model of such a system, as described in section 2.1.1. Ideally the training of these models will involve data that is *labelled*, i.e. for each recording the corpus will provide an annotation that states the time indices at which each phoneme contained in the audio begins and ends, thereby indicating the exact boundaries of each individual phonetic unit. This approach is referred to as *supervised training*. The necessary labels usually have to be generated via time-consuming manual annotation of a recording by linguists. Alternatively an existing system for the target language can be used in an attempt to annotate further audio by performing a so-called *forced alignment*, as suggested in [BrFO93]. Forced alignment is an HMM based technique that uses the Viterbi algorithm to align the (pre-processed) speech signal to the known phonetic sequence (i.e. the aforementioned label sequence). Analogous to training on labelled data, this form of segmentation is referred to as *supervised segmentation*. Conversely, *unsupervised segmentation* does not utilize any prior knowledge about the structure or contents of an utterance. Instead, an unsupervised system will look solely at the

(again pre-processed) audio and attempt to segment it based on that information alone, according to its acoustic models (which may have been trained in a supervised or unsupervised fashion). The segmenters presented in this work fall into the unsupervised category, although they have been *trained* using existing labels, i.e. in a supervised manner.

It should be noted that the segments resulting from a cross-lingual segmentation approach such as the one used here (and the subsequent clustering of the derived units, which is outside the scope of this work) may not correspond to the acoustic units human experts would use to classify the phonetic inventory of the targeted language. The segmentation may be finer (with otherwise atomic segments further sub-divided) or coarser, leading to a larger or smaller number of total segments respectively. While it would be ideal to automatically generate an inventory which is congruent with that obtained through manual analysis by linguists, this is, in practice, not realistic.

Section 2.3 below will present an overview of previous efforts on the subject of unsupervised phoneme segmentation.

2.3 Related Work

Significant work has been done on the topic of building speech recognition systems for phoneme segmentation on unwritten and under-resourced languages.

Inspiration for our approach was taken from experiments conducted by Muthukumar and Black in [MuBl14] on the automatic discovery of phonetic inventories without prior phonetic knowledge about the target language. The authors used a cross-lingual, neural network based articulatory feature predictor in combination with hierarchichal clustering in order to construct a phoneme set for use in speech synthesis.

In [ScWE09] Scharenborg et al. attempted to hypothesize phoneme boundaries based on acoustic change in the audio signal, and compared these estimated boundaries to those created manually by human transcribers. Identified errors were found to be related to ”segment duration, sequences of similar segments, and inherently dynamic phones”. They proposed to expand existing one-step methods to two-step approaches, mixing commonly used bottom-up information taken from the signal with top-down information.

Kuo et al. also suggest a two-stage approach in [wKLW07] that attempts to mimic the human phoneme segmentation process. In a first step they use Hidden Markov Models to perform forced alignments according to a ”minimum boundary error criterion”. The second stage employs Support Vector Machines for refinement.

In [yLGl12] Lee et al. approached the issue of acoustic modelling in the absence of both, pre-knowledge about the target language and annotated training data. They employed a Dirichlet process mixture model to represent sub-word units, allowing them to simultaneously segment the speech signal and discover a phonetic inventory for the target language, complete with Hidden Markov Models for all discovered acoustic units.

In [QiSM08] the authors attempted to develop a series of objective functions in order to determine segmentation quality. They then employed a time-constrained

agglomerative clustering algorithm to minimize these objective functions and arrive at an optimal segmentation. They further improved their results in [QiMi08] by introducing Minimum of Summation Variance and Maximum of Discrimination Variance in order to determine parameters for optimizing segmentations according to the Mahalanobi distance metric.

Recently work on the subject has been done as part of the Zero Speech challenge, as presented in[VTSC+15]. The challenge is centered around the unsupervised discovery of subword units from raw speech, providing a "unified and open source suite of evaluation metrics and data sets" to make results of various approaches comparable and facilitate analysis.

Estevan et al. have applied Maximum Margin Clustering (MMC) to the task of segmenting speech on a phonetic level in [EsWS07]. A kernel method, MMC is a (semi-)unsupervised form of SVM that uses the maximum margin criterion to find an optimal solution for the segmentation. Results were evaluated using correct detection rate, over-segmentation and a false alarm rate indicating the relative frequency with which boundaries are incorrectly detected.

In [AEEM01] and [EsAv04] Esposito and Aversano attempted to use sharp transitions, or spectral instability, of the short-term transform of speech signals to implement a bottom-up approach that allows segmentation of recorded speech independent from language, linguistic context or a written representation.

3. Experimental Setup

In this chapter we will briefly introduce JANUS, the recognition toolkit used in our experiments, as well as the employed corpora and data sets. We will also look at the phoneme coverage of the chosen source languages on the target language, which we will later use in order to investigate the correlation between phonetic similarity and performance. Finally, we will give a brief overview of how our mono- and multilingual ASR systems were trained.

3.1 JANUS

The speech recognition systems used for the experiments presented here were trained and applied using the JANUS Recognition Toolkit (JRTk) [LWLF+97]. The JRTk was developed in a co-operation between the Karlsruhe Institute of Technology and Carnegie Mellon University. Functional modules are accessed via an object oriented approach using a Tcl/Tk scripting interface. JRTk features the IBIS decoder, allowing for single-pass decoding, and uses Hidden Markov Models for acoustic modelling [SMFW01].

3.2 Data

Since we are working under the assumption that there will be no annotated data available for our intended target language in the actual use case (either due to lack of a writing system or the cost associated with generating such annotations), we will need to train the recognizers that are going to be used for segmentation on data taken from other languages. Furthermore, in order to investigate and compare the characteristics of the produced segmentations, it would be preferable to test the approach on various sets of target data, ideally across different languages.

The following sections will introduce and describe the data sets used in the experiments as source and target audio.

3.2.1 Training data

For our experiments we chose German (DE), French (FR), Italian (IT), Turkish (TR) and Russian (RU) as source languages for both, mono- and multilingual systems. The audio and annotations were taken from the Euronews corpus [Gret14]. Euronews is a collection of news recordings from the multilingual television station of the same name. The data was collected in a manner so that audio is available from each language for every chosen news event. Due to the nature of the recordings, the speech is often superimposed over a video report that features separate audio, which means the audio must be considered overall noisy.

For each of the chosen source languages we used a subset of the Euronews training data consisting of approximately 70 hours of audio, on which individual, monolingual recognizers were trained. The English recognizer serving as a gold standard was also trained on 70 hours of Euronews data. Models for the multilingual system (referred to as M5) were estimated using a combined training set of roughly 360 hours of German, French, Italian, Turkish and Russian audio. For precise numbers please see table 3.1.

Language	EN	DE	FR	IT	TR	RU	M5
Length	72.8h	73.2h	68.1h	77.2h	70.4h	72.2h	361.3h

Table 3.1: Amount of audio data used for training mono- and multilingual recognizers

3.2.2 Test data

The following sections will give a short description of the test data used to evaluate the quality of produced segmentations.

3.2.2.1 English test data

For segmentation tests on English we used data from two different corpora: Euronews and TIMIT [Garo$^+$93]. The Euronews data used for testing consisted of a separate set of news broadcasts with an approximate length of 4 hours total. Like the training data it is characterized by the actual speech being superimposed over the background audio of a news report, and must therefore be considered noisy.

TIMIT, on the other hand, is a corpus consisting of speech recorded expressly for the purpose of developing and evaluating ASR systems. It consists of individual recorded phrases spoken by 630 speakers of eight different American English dialects. Each speaker was asked to speak 10 different sentences that are meant to represent the overall phonetic characteristics of the English language in a controlled, noise-free recording environment. The corpus offers time-aligned annotations not only on an orthographic, but also on a phonetic level, which makes it uniquely advantageous for the purpose of evaluating the accuracy of an automated phoneme segmentation.

It should be noted that with these characteristics the audio differs in two major aspects from that contained in the Euronews training and test sets. Firstly, it represents a different accent of the English Language (American, rather than English). And secondly it is clean rather than noisy. We will look at whether these differences affect the performance of our process in chapter 4, Evaluation.

3.2.2.2 Basaa test data

Basaa is one of three Bantu languages used in the BULB project, and therefore an ideal target language for initial experiments on phoneme segmentation. As of 2005 there were approximately 300,000 live speakers of Basaa across several regions in southern Cameroon [LeSF15].

The data used in our experiments consisted of roughly 2 hours of re-spoken radio broadcasts, i.e. the original wording was transcribed and then re-recorded in a quiet environment in order to achieve overall higher quality recordings. While the original speech was that of a male speaker, the re-speaking was performed by a female native speaker of the language using a voice-memo application.

3.3 Phoneme Coverage

The phoneme sets of employed source languages will differ both in terms of size as well as coverage on a target language. For English a set of 40 phonemes was used in training, whereas the phoneme sets of individual monolingual systems for the chosen source languages ranged between 25 and 59 phonemes. Since all dictionaries were created via the G2P component of the Mary text-to-speech system ([ScTr03]) using the same phone set, no mappings were required to merge the phone sets of the individual languages for the multilingual segmenter, which was trained using a combined set of 99 distinct phonemes.

Intuitively one would assume that a phonetic similarity between source and target languages would positively influence the cross-lingual performance of a recognizer that operates purely on a phonetic level. In order to check for a correlation between phonetic similarity and segmentation quality we calculated phoneme coverages for the monolingual as well as the multilingual recognizers. The results are presented in table 3.2. We will later attempt to correlate these percentages with the performance scores achieved by their respective systems once we have obtained results.

Language	#Phonemes	% Coverage
EN	40	–
M5	**99**	**85.0**
DE	56	82.5
FR	33	57.5
IT	59	60.0
TR	26	55.0
RU	25	37.5

Table 3.2: Number of phonemes used in training and phoneme coverage on English

3.4 Training of monolingual and multilingual acoustic models

The following sections will give a brief overview of the training of both mono- and multilingual acoustic models.

3.4.1 Monolingual acoustic model training

Training data for monolingual systems consisted of 16KHz audio recordings, which were used to train HMM models for all phonemes contained in the phoneme sets of each respective source language, individually. We used a standard three-state subphoneme model with begin, middle and end states for each phoneme and a simple topology that allowed for transitions to be either recurrent or jump forward one state. To optimize robustness of models across language barriers, and to make results comparable to potential future gold standard segmentations using target languages that do not offer large amounts of training data, we trained context-independent acoustic models in all cases. Since the training data provided by the Euronews corpus does not have transcriptions on a phonetic level, a flat start was performed, followed by six more iterations of training, using feature vector stacks consisting of 13 Mel-frequency cepstral coefficients stacked over seven time indices left and right, with a window size of 16ms and a frame shift of 10ms.

3.4.2 Multilingual acoustic model training

Multilingual models used the same process, parameters and topologies for training to ensure comparability to our monolingual models. The central question of multilingual training is the manner in which the information contained in the training data of the individual source languages is combined. In [ScWa01] the authors have proposed three different ways of combining HMM-based acoustic models for the purpose of language-independent acoustic modeling, which is sufficiently similar to our task of cross-lingually applying multilingual models to unseen languages. Separate acoustic modeling involves no sharing of data for the actual training of HMM parameters, but is limited to a shared LDA matrix calculated based on all language-specific phoneme models. Alternatively, (sub-)phonemes are tagged as belonging to their respective languages (e.g. M-b/DE), and the gaussians that represent the emission probabilities of the HMM are shared, while the mixture weights are not. Finally, a truly mixed multilingual model shares both gaussians and mixture weights, reducing the multilingual phoneme set to the simple set union of the language-specific sets. For our purpose we chose the latter approach, combining the available audio of all source languages into one large corpus, while merging pronunciation dictionaries, and phoneme sets accordingly.

4. Evaluation

Once a phoneme segmentation for a specific set of test data has been created by the previously trained system, the quality of that segmentation needs to be determined. In this chapter we will first introduce the metrics chosen for this assessment and then present the results that were achieved with several recognizers on various sets of test data. We will also investigate how the use of language models impacts performance, and if there is a correlation between segmentation quality and the phoneme coverage of the source language(s) on the target language.

Finally, we will look at the aspect of oversegmentation, i.e. the ratio between the number of produced segments and the actual number of segments, as provided by the ground truth.

4.1 Metrics

In order to evaluate the segmentations obtained with a trained recognizer, we require an objective metric to measure the correctness of the predicted segment boundaries.

A first approach to this effect was to calculate the distance between each predicted boundary and the closest segmentation index in the ground truth and take the average of these distances. However, the choice of parameter values in decoding heavily influences the number of segments generated in the hypotheses. Since a higher number of segments, and therefore a higher number of boundaries, will always decrease the average distance between predicted and true segmentation indices, this approach turned out not to be useful for our purposes. The same logic inversely applies to the use of oversegmentation, i.e. the ratio between the number of hypothesized segments and the number of actual segments in the ground truth - the more segments are predicted, the worse an oversegmentation score is going to be, therefore favouring parameter settings that produce segmentations with *fewer* boundaries. This issue can be solved by normalizing oversegmentation such that it becomes an over-/undersegmentation metric that is centered around 0, and evaluating a segmentation by how far it deviates from the ideal value of 0. However this still only takes into account the total number of segments generated, while ignoring whether the location (i.e. the time index) of a boundary is anywhere close to one in the ground truth.

We next turned to the Correct Detection Rate (CDR) as introduced in [ScWE09]. Scharenborg et al. define CDR as follows:

$$CDR = \frac{\#boundaries\_correct}{\#boundaries\_truth}$$

This metric suffers from the same issues presented before. In the case of CDR, since the number of *all* hypothesized segments is not a factor, the score will be more favourable the more boundaries are hypotesized, since this will make matching a true boundary more likely. However, CDR as defined above is fundamentally no different from the recall metric commonly used in the evaluation of pattern recognition systems.

It therefore seemed logical to combine CDR (from here on referred to as recall) with precision in order to evaluate our segmentations. For this we define precision as

$$precision = \frac{\#correct\ detections}{\#segments\ in\ hypotheses}$$

and recall as

$$recall = \frac{\#correct\ detections}{\#segments\ in\ reference}$$

That is, our precision values show what percentage of our predicted boundaries are correct, while our recall shows what percentage of all *actual* boundaries - as indicated by the ground truth - has been found.

Inverse to the monotonic increasing behaviour of recall, precision scores will *decrease* the more boundaries have been hypothesized, due to the absolute number of predicted segments present in the denominator of the equation.

The F-measure takes into account the generally opposed behaviour of precision and recall to combine them into a weighted average with resulting values ranging between 0 and 1. For our purposes we choose the F_1-score, which incorporates precision and recall with equal weights, therefore representing the harmonic mean of the two values:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

In order to evaluate to what extent the determined phoneme coverage of a source language impacts the respective monolingual system's performance, we are going to calculate the *Pearson correlation coefficient* (PCC) as a measure of the linear correlation between coverage and precision, recall and F_1-score, respectively. It is defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

with X and Y representing coverage and the chosen score, *cov* the covariance and σ the standard deviation. PCC values range from 1.0 (total positive correlation) to -1.0 (total negative correlation), with 0 indicating no correlation one way or the other.

4.1.1 Tolerance

Without further modifications, in order for a predicted phoneme boundary to be labeled as "correct" its time index would have to exactly match that of a ground truth boundary. Due to the ambiguity of speech signals, reaching absolute precision when predicting the start and end of a phonetic unit is unrealistic. We therefore introduce an error *tolerance* when applying the metrics presented in section 4.1.

The recognizers used in the experiments presented here work with a window shift of 10 ms. I.e. the time indices of all segment boundaries, both true and hypothesized, are always multiples of these 10 ms. We opted to allow for an inaccuracy of 2 indices to both sides of a reference boundary when deciding whether a predicted boundary should be considered correct or not. This equates to a tolerance of 20 ms, which is a value that has also been used in the past by other researchers, such as [ScWE09] or [QiSM08]. Higher tolerances of 3 or more indices (i.e. 30 ms or higher) were also applied, but quickly escalated the values received for our metrics. Therefore all results given here are for a tolerance of 20 ms, without further indication.

4.2 Performance results

4.2.1 Performance without language models

The following results were achieved without the use of language models. Instead, all phonemes in our vocabulary were marked as noise and a noise penalty was applied in place of a LM score.

4.2.1.1 Performance on Euronews English

In initial experiments we employed English as a *faux* unseen target language. I.e. we ran phoneme decodings using recognizers previously trained on source languages other than English on English audio, assuming no prior knowledge of the language's structure, vocabulary or phonetic inventory.

One set of target audio was taken from the same corpus as the training data, Euronews. The primary issue with this data is that Euronews does not provide annotations on a phoneme level, and as such offers no manually created ground truth for our evaluation. We therefore decided to also train a recognizer on the separate English training data, create a phonetic annotation from the orthographic one provided by the corpus (using the G2P component of the Mary Text-to-Speech system), and then run a forced alignment between the target audio and this new phonetic annotation. This process resulted in a time-indexed sequence of phonemes

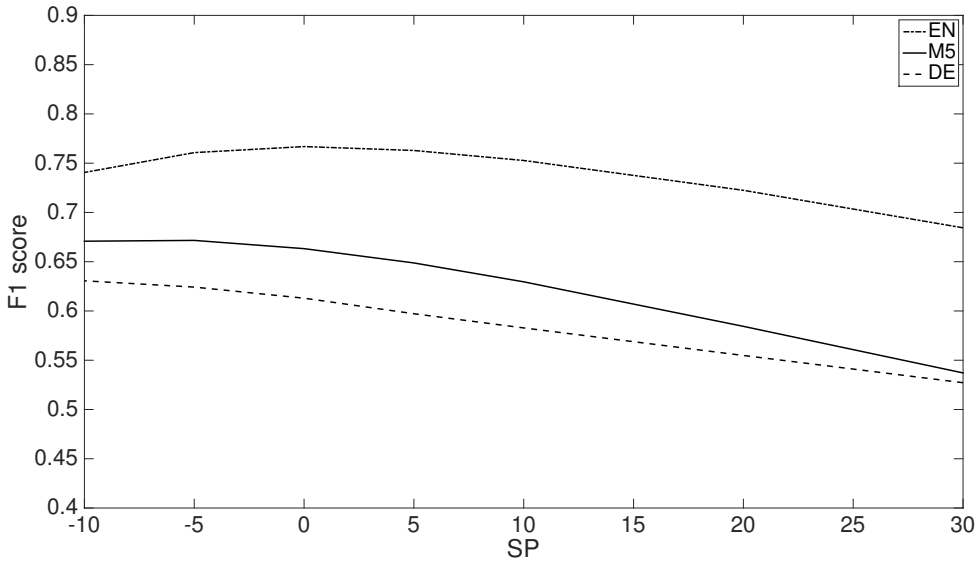

Figure 4.1: F_1-scores for segmentations on English Euronews audio

which could then be used to evaluate the accuracy of our hypothesized segment boundaries.

For comparing results on this data we chose three different recognizers: a single-language cross-lingual one, the multilingual system trained on all five chosen source languages, and a recognizer trained on the actual target language to represent a baseline. For the monolingual system we picked a recognizer trained on German audio, as we expected this system to perform best given that it showed the highest phonetic similarity to the target language, English.

Figure 4.1 displays the F_1-scores calculated for the decodings with all three systems. The x-axis shows the chosen values for the silence penalty (SP), which substitutes for the word penalty (LP) when not using a language model. Higher SP values will result in fewer segments in the resulting hypotheses, and vice versa. Figures 4.2 and 4.3 show the according precision and recall scores on the same data, respectively.

Figure 4.3 shows that recall drops continuously for increasingly higher SP settings - the fewer segments are generated in a decoding, the lower the chance of matching a ground truth boundary within the given tolerance. Conversely, if (in theory) the silence penalty (or word penalty, when using a language model) were set low enough, the system could potentially hypothesize boundaries at every possible time index and thereby easily match every true boundary in the reference, having the recall approach a value of 1.

As is to be expected, precision behaves exactly opposite. A value of 1 could only theoretically be achieved with an oversegmentation of precisely 0, i.e. when the hypothesized segmentation contains exactly as many segments as the reference. Therefore the more segments are produced (at lower silence penalties), the more precision will drop.

Both precision and recall already show the expected pattern when it comes to comparing the performance of all three chosen system: the recognizer trained on the

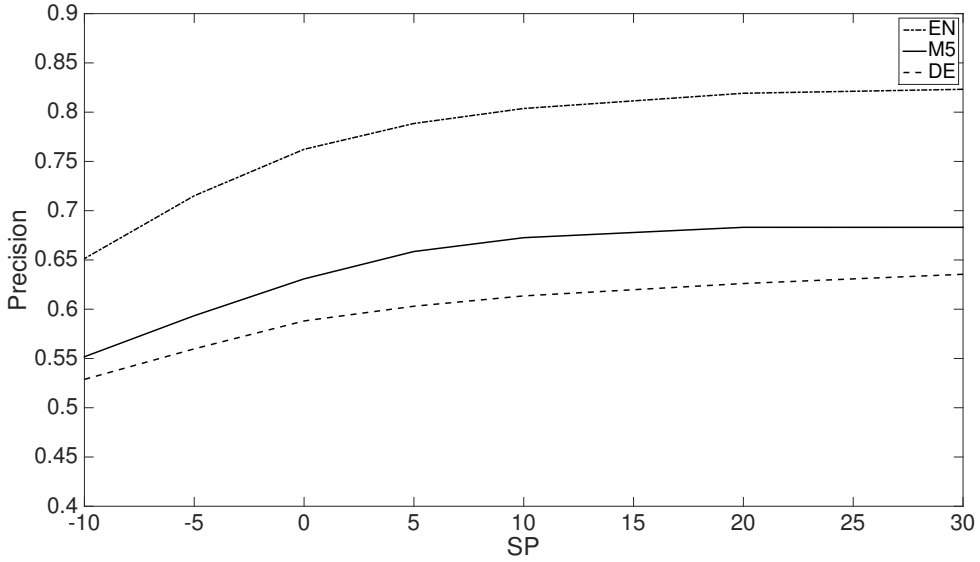

Figure 4.2: Precision scores for segmentations on English Euronews audio

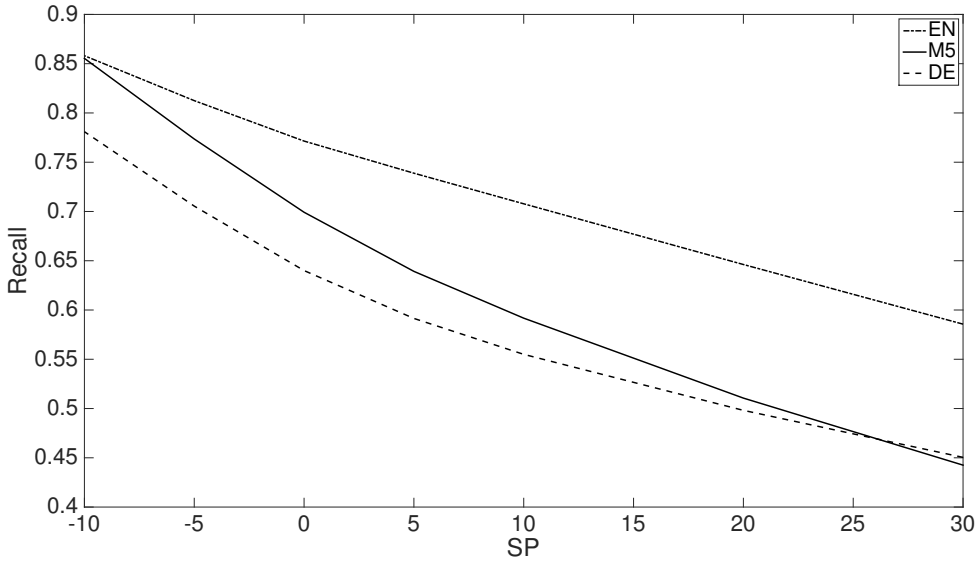

Figure 4.3: Recall scores for segmentations on English Euronews audio

target language as a baseline clearly outperforms both mono- and multilingual cross-language systems, while the multilingual system scored higher than the monolingual one throughout.

Consequently, this behaviour is also reflected in the F_1-scores, ranking the performance of the baseline system first, followed by the M5 and German recognizers. What differs is the parameter setting at which each of the systems achieves its best result according to the metric. The recognizer trained on English audio performs best at a silence penalty of 0, while the multilingual system seems to plateau between values of -10 and -5, with a possible peak somewhere in between. The curve for the German recognizer seems to continue to grow for SP values below -10. However, we did not run further tests with even lower values, as the number of produced segments would grow so far beyond the actual number of segments in the audio that the result could hardly be considered useful. Table 4.1 shows the F_1-scores of all three systems for SP values around their respective peaks.

SP	-10	-5	0	5
EN	0.7406	0.7607	**0.7669**	0.7629
M5	0.6708	**0.6716**	0.6633	0.6487
DE	**0.5769**	0.5429	0.4982	0.4510

Table 4.1: F_1-scores for segmentations on English Euronews audio

4.2.1.2 Performance on TIMIT English

Once experiments on the English Euronews test data were concluded, we applied the same approach to a set of audio taken from the TIMIT corpus. Unlike the tests performed on the English target audio taken from Euronews, evaluating the segmentation of the TIMIT test data did not require creating annotations on a phonetic level via decodings and forced alignments, since the TIMIT corpus already provides manually created phonetic transcriptions, complete with time indices. The evaluation was performed using the same three systems used on the English euronews data, as described in section 4.2.1.1. Results for all three systems can be found in figures 4.4, 4.5 and 4.6.

As shown in figure 4.6, the recall curves seem to exhibit a similar behaviour as previously seen in the results on Euronews, with a slightly weaker overall performance which could be attributed to a less perfect match between the noisy training data and the very clean audio provided in the TIMIT corpus. However, the relative performance between the three systems is rather unexpected. The multilingual recognizer performed best with regard to recall, followed by the monolingual one, with the English recognizer scoring visibly worse, especially at lower SP values. This is in contrast to the expected order that was observed previously, with EN followed by M5 and then DE.

Figure 4.5 shows that precision behaves even more erratic. The ”curve” is almost completely flat and shows virtually no response to the chosen SP values. Although for most of the applied SP values the expected relative order between the systems seems to hold, the extremely small absolute differences in score suggest that this order may as well be due to random variance. Since the F_1-score represents an

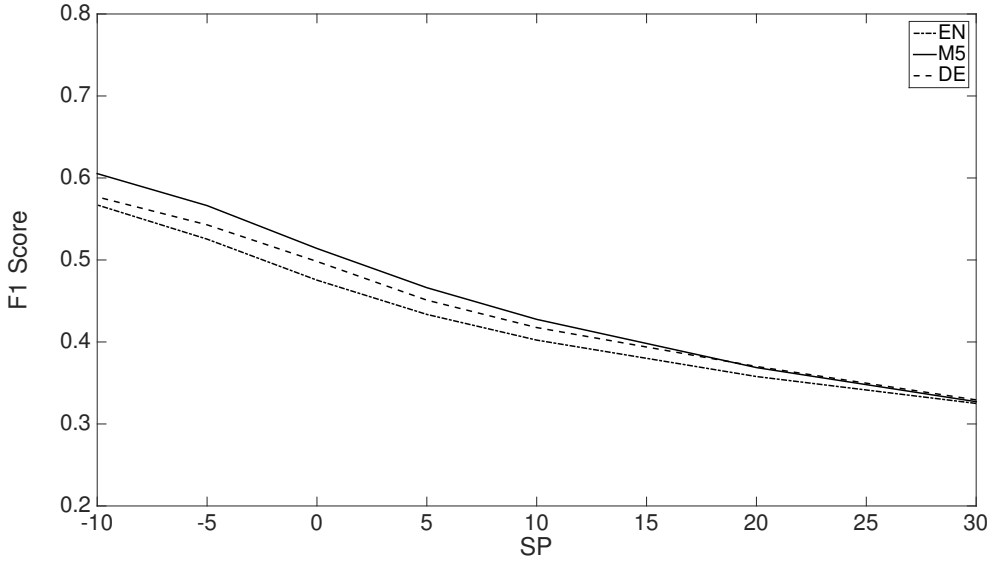

Figure 4.4: F_1-scores for segmentations on English TIMIT audio

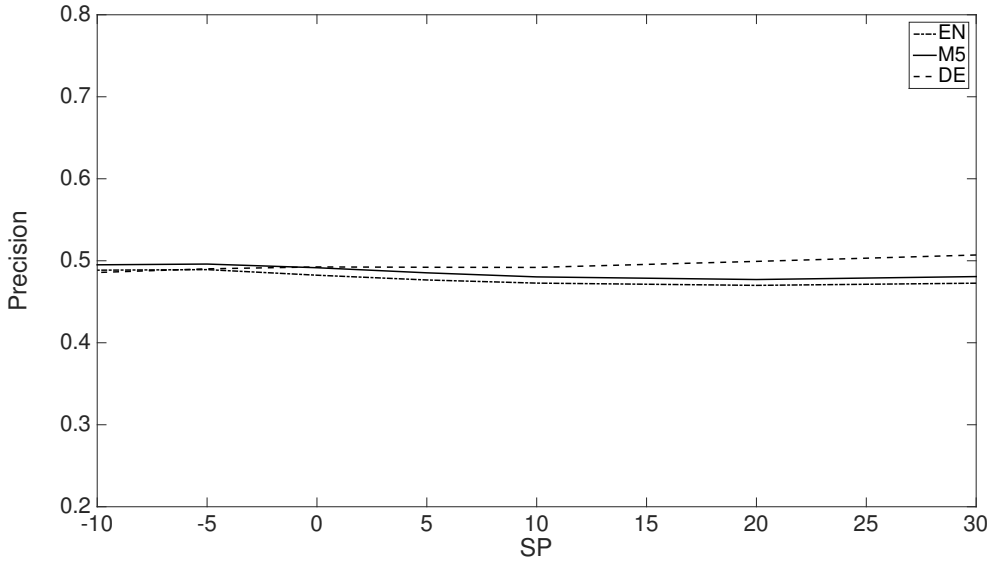

Figure 4.5: Precision scores for segmentations on English TIMIT audio

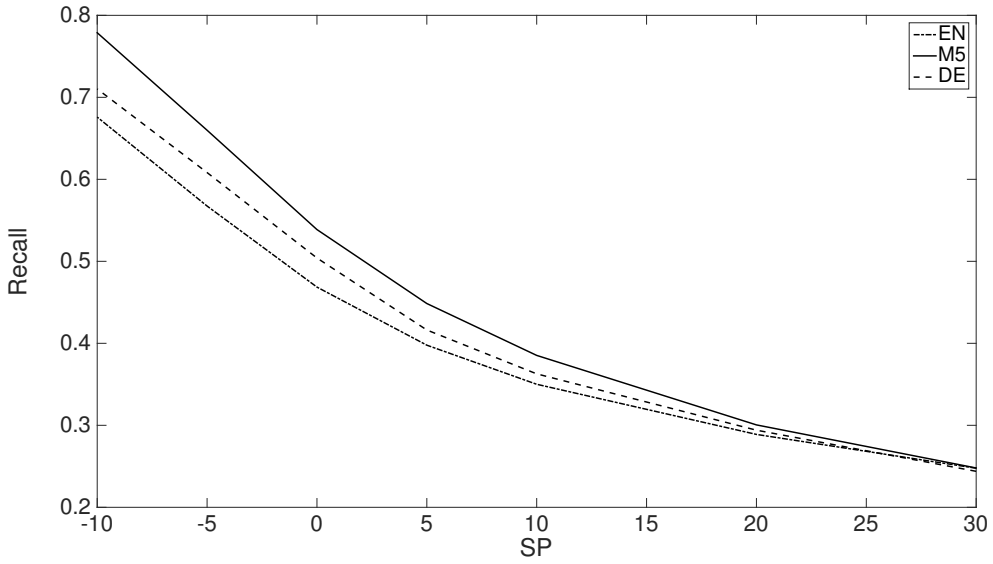

Figure 4.6: Recall scores for segmentations on English TIMIT audio

equally weighted average of precision and recall, the quasi-flat precision curve results in a behaviour of the F_1-curve that very closely mirrors that of the recall.

Were these results to be taken at face value, the conclusion would be to set system parameters in a way that produces as many segments as possible, thereby maximizing recall and consequently the final metric. This is obviously erroneous, which means the results on TIMIT audio must be discarded as faulty.

From a purely technical point of view we should have expected the segmentations to behave similarly to what we have observed on the English Euronews data. We could not find any errors in the implementation or the data itself. The major differences between the two data sets were the noisiness of the audio and the accents of the speakers. As previously described in section 3.2.2.1, Euronews consists of speech that is often superimposed over news footage containing separate audio, while TIMIT speech was recorded in a controlled environment without any background noise. Also, Euronews English audio usually features British speakers, whereas TIMIT was recorded by speakers representing several different American accents.

Maybe these mismatches contributed to the unpredictable behaviour when attempting to conduct decodings on a phonetic level, although the extent to which this seems to impact performance seems much bigger than what one would expect. Furthermore, as we will see in section 4.2.1.3, similar experiments on Basaa audio did not exhibit the same flat, quasi-random precision curves. Since the re-spoken African data was also devoid of noise and the phonetic differences between English and Basaa are obviously more significant than those between American English and British English, the observed behaviour can likely not be attributed to these factors.

4.2.1.3 Performance on Basaa

Finally, we applied all source language systems (mono- and multilingual) to the available Basaa data. Since our experiments using special language models (estimated on phoneme sequences taken from the training data's annotations) did not result in

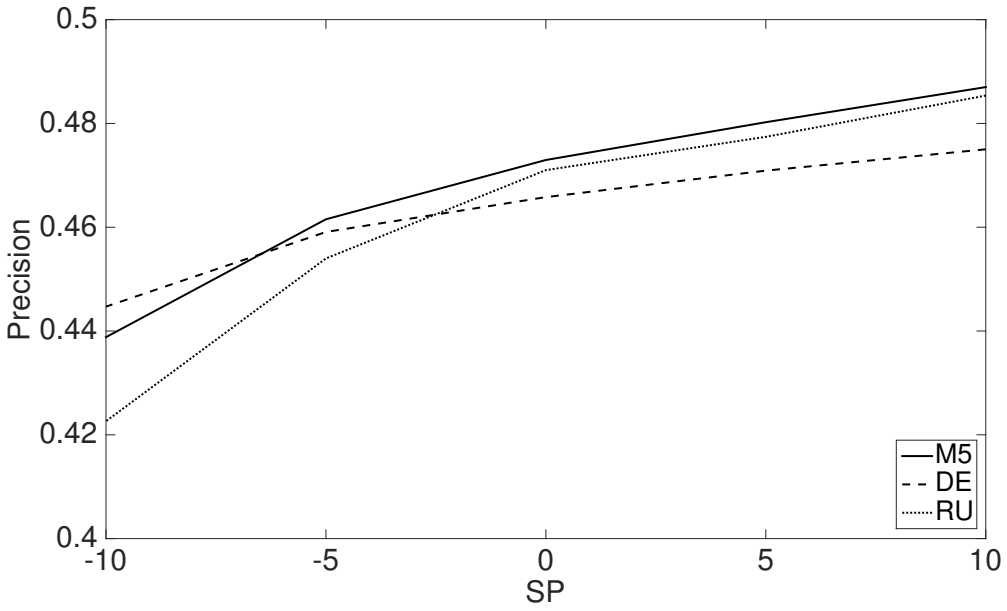

Figure 4.7: Precision scores for segmentations on Basaa audio

a significant improvement (see section 4.2.2), we returned to working without such models for these tests. For the sake of readability not all monolingual systems have their results represented in figures 4.7, 4.8 and 4.9. Instead we chose to display only the multilingual results, as well as the best- and worst-performing source languages (as per F_1-scores), which are Russian and German, respectively. Full results can be found in tables 4.2, 4.3 and 4.4.

Note that for tests on Basaa there is no gold standard to compare against, since the authors were not aware of any published work on this particular language at the time of writing, nor did they have access to a recognizer trained on it. We therefore present the results as they are, without comparison to a previously established baseline.

SP	-10	-5	0	5	10
M5	0.4388	0.4615	0.4730	0.4802	0.4870
DE	0.4447	0.4591	0.4658	0.4709	0.4750
FR	0.4657	0.4966	0.5166	0.5274	0.5331
IT	0.4519	0.4689	0.4808	0.4901	0.4982
TR	0.4642	0.4787	0.4891	0.4952	0.5007
RU	0.4226	0.4540	0.4710	0.4774	0.4854

Table 4.2: Comparison of Precision on Basaa audio (without language models)

Unlike the previous decodings on TIMIT audio (see section 4.2.1.2), the Precision, Recall and F_1-curves do behave mostly as expected. Recall increases with lower penalty values (i.e. with more estimated segments), Precision with higher penalty values (i.e. with fewer estimated segments). Although the curve (shown in Fig. 4.7) is flatter than that for our tests on Euronews English (cp. section 4.2), it does not display either the complete flatness of the precision curve for TIMIT or its randomness with regard to how the different systems compare amongst each other.

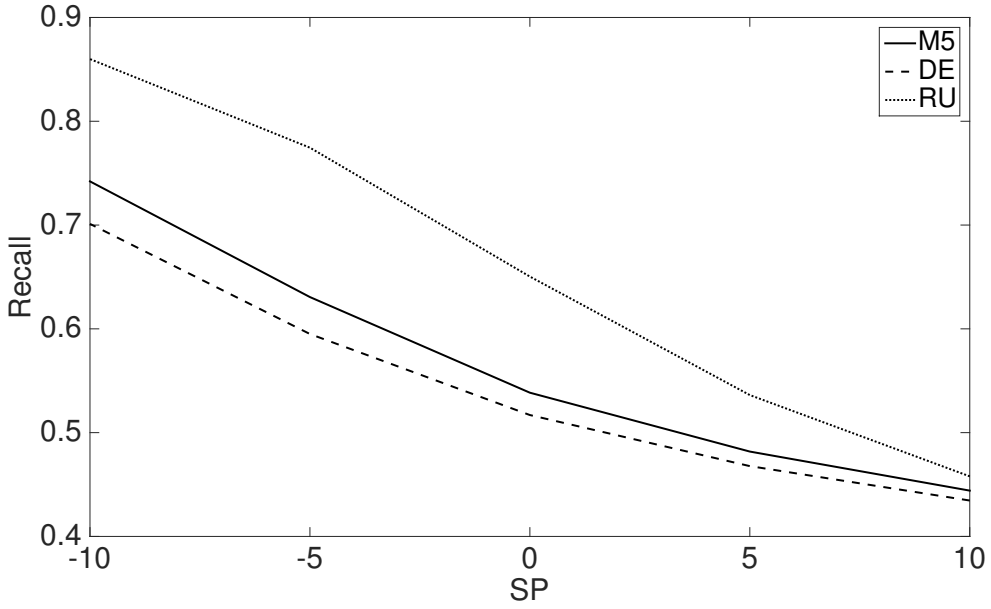

Figure 4.8: Recall scores for segmentations on Basaa audio

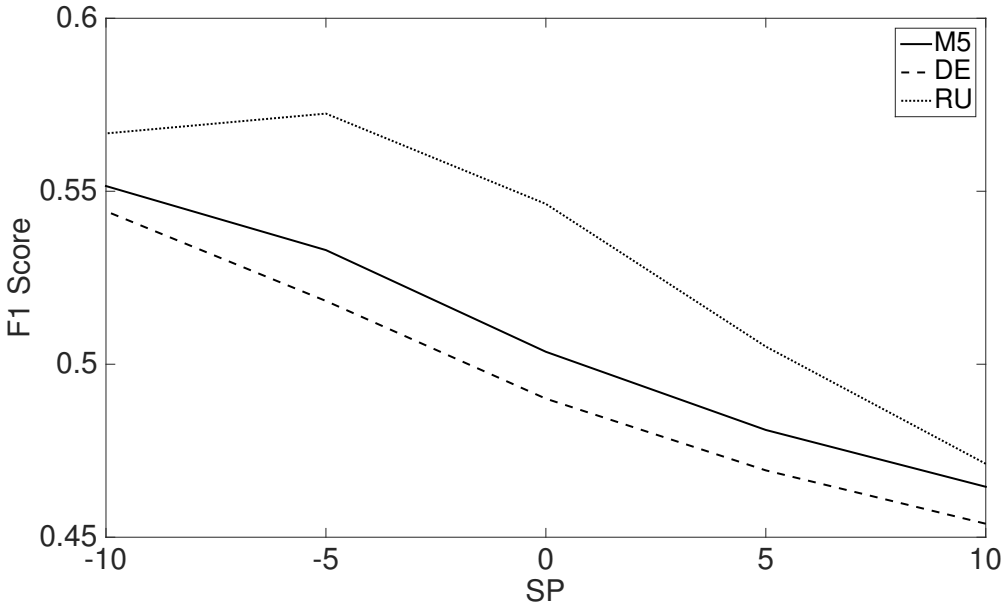

Figure 4.9: F_1-scores for segmentations on Basaa audio

SP	-10	-5	0	5	10
M5	0.7421	0.6306	0.5385	0.4818	0.4441
DE	0.7011	0.5950	0.5170	0.4677	0.4346
FR	0.6989	0.5881	0.5149	0.4711	0.4413
IT	0.7065	0.5997	0.5209	0.4737	0.4435
TR	0.7479	0.6426	0.5538	0.4917	0.4504
RU	0.8599	0.7745	0.6504	0.5362	0.4579

Table 4.3: Comparison of Recall on Basaa audio (without language models)

SP	-10	-5	0	5	10
M5	0.5515	0.5330	0.5036	0.4810	0.4646
DE	0.5442	0.5183	0.4900	0.4693	0.4539
FR	0.5590	0.5385	0.5158	0.4977	0.4829
IT	0.5512	0.5263	0.5001	0.4818	0.4692
TR	0.5728	0.5487	0.5195	0.4934	0.4742
RU	0.5667	0.5724	0.5463	0.5051	0.4712

Table 4.4: Comparison of F_1-score scores on Basaa audio (without language models)

Just like the TIMIT audio, the Basaa data used here has been recorded in a quiet environment without background noise, however the resulting F_1-curve shown in figure 4.9 does not as strongly mirror the shape of the underlying recall. We must conclude that the irregular behaviour on the TIMIT audio does not (entirely) stem from the audio mismatch between the noisy training and the clean test data, or the difference between the American and British accents (as Basaa certainly is further removed from the training data, acoustically, than a mere accent).

For Russian, the best-performing single language, there is a peak at SP -5. German and the multilingual segmenter do not show a peak within the chosen parameters; it is likely the optimum lies further towards lower SP values, since in our experiments the Russian decoder has generally displayed a tendency to produce more segments than other systems for the same parameter settings. Since optimal results can be expected with an oversegmentation close to 0 (i.e. for a number of predicted segments roughly equal to those found in the ground truth), curves for Russian will be be shifted to some extent. The reason for this behaviour is not obvious; it may very well stem from some fundamental phonetic characteristics of the Russian language. Further investigation would be required in order to ascertain if this is the case.

4.2.2 Performance using language models

In addition to the experiments without language models presented in section 4.2.1, we also ran tests using special language models, incorporating new information about the acoustic properties of our source languages.

As described in section 2.1.1, the acoustic model represents knowledge about the individual phonetic units (and sub-units) of a language. While in case of context-dependent systems these models do take into account limited amounts of left and

right context, the context-independent models we used in our experiments (for the reasons given in section 3.4) did not. For this reason we introduced language models to our experiments as an additional source of information that can take into account the probabilities of specific successions of words over a chosen context. For the purpose of phone segmentation the vocabulary of the language model will consist of the individual phonemes of the language, rather than the usual words.

For initial tests we decided to use simple trigram language models. Since the Euronews corpus does not provide annotations on a phonetic level, we generated them using the G2P component of the MARY text-to-speech toolkit. We then estimated a trigram model with Witten-Bell smoothing for each individual source language, as well as a mixed model trained on data taken from all five. Finally, we re-ran some of our previous experiments, now using the estimated probabilities of the language models.

The first observation we could make was that segmenting the audio was now considerably slower than before. Whereas decoding without language models was performed in approximately real time, the same task using language models would take up to 200 times real time. Therefore, in order to conduct language model experiments in an at least somewhat reasonable time frame, we decided to reduce the amount of test data by removing the longest individual utterances from the sets, since each provided utterance is decoded by a single process without further parallelization. This reduction of the test data cut decoding time by two thirds while only reducing the amount of test data by approximately 17% (from 29 minutes to 24 minutes total).

As for performance, initial results showed a marginal improvement in F_1-score of approx. 1% absolute for the multilingual system (using a mixed multilingual phoneme language model). Of the monolingual segmenters we chose to run these decodings with German only, due to the considerable time expense. The German decoder (using a German monolingual phoneme language model) increased performance slightly more than the M5 one, with an improvement in F_1-score of approx. 3% absolute. However the English baseline system actually performed *worse* using a monolingual English phoneme language model, losing approx. 3% absolute. Considering how unreliable the effect of applying these language models is, and that performance gains, where present, are marginal at best, it would be difficult to justify the considerably longer decoding times. A summary of the results can be found in table 4.5.

System	without LM	with LM
EN (baseline)	0.7769	0.7484
M5	**0.6624**	**0.6708**
DE	0.6130	0.6446
M5*	n/a	0.5908

Table 4.5: Comparison of segmentation results with and without language models

We suspected that the severe increase in computational expense might be caused by the size of the lattice generated for the language model search. Since language models are usually trained and applied on a word basis, applying the same parameters to

a lattice generated for phonemes may lead to very large search spaces due to the increased number of units per utterance. To test for this we modified the parameters regulating the lattice generation and search beam. While these restrictions *did* lead to much faster decodings, bringing the time expense back to roughly real time, performance suffered significantly from these changes, with the multilingual system losing 7% absolute compared to the original tests without language models, which ran equally as fast.

4.3 Other aspects

4.3.1 Phoneme coverage and performance

Our initial assumption was that phonetic similarity between source and target languages should positively influence the quality of a segmentation. One way to evaluate this similarity is to calculate a phoneme coverage for each source-target pair, i.e. what percentage of the target language's phoneme set is also part of the set used when training the system on the source language(s).

As shown in table 4.6, the highest single-language coverage on English was achieved by the german system with 82.5%. The lowest coverage is that of the Russian phoneme set, at 37.5%. The multilingual set, again listed as "M5", has a marginally higher coverage than German at 85%.

Language	F_1-score	Precision	Recall	% Coverage
EN	0.7669	0.7623	0.7715	100.0
M5	0.6624	0.6299	0.6984	85.0
DE	0.6130	0.5881	0.6400	82.5
FR	0.6748	0.6748	0.6706	57.5
IT	0.6631	0.6286	0.7016	60.0
TR	0.6449	0.6208	0.6709	55.0
RU	0.6338	0.5559	0.7372	37.5

Table 4.6: Phoneme coverages and scores for English as target language

We then take the results for the evaluation metrics discussed above (Precision, Recall and F-Score) and calculate the Pearson correlation coefficient. The results can be found in table 4.7.

Metric	PCC
F-Score	0.5620
Precision	0.6518
Recall	0.1494

Table 4.7: Correlation between phoneme coverage and evaluation metrics on English

As we can see, there is no universally strong correlation between phoneme coverage and our chosen metrics. At 0.1494, the PCC for recall indicates close to no correlation, while the coefficient for precision is somewhat significant at 0.6518. Our chosen

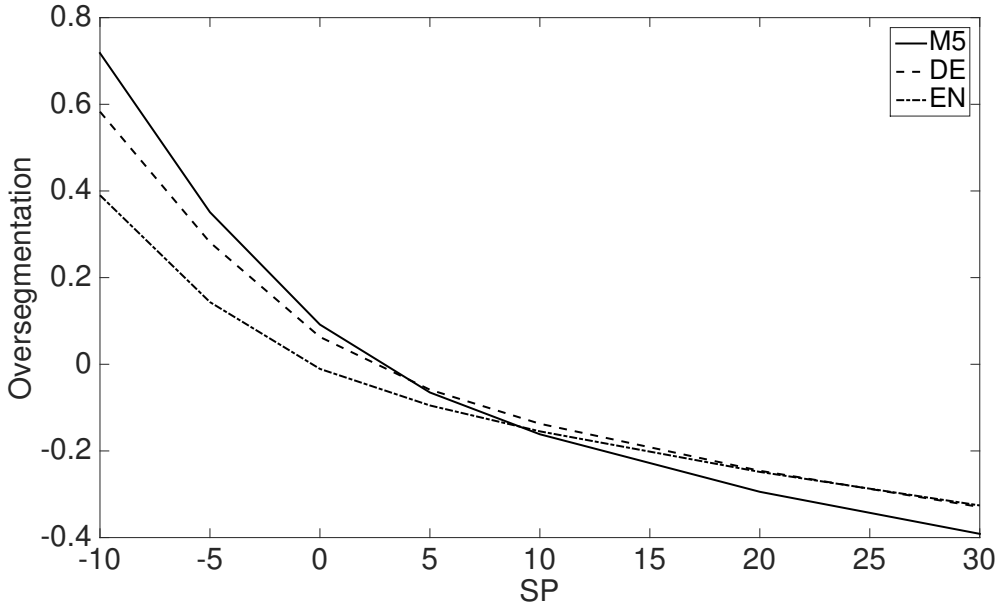

Figure 4.10: Oversegmentation of 2 segmenters plus baseline on English Euronews audio

primary metric, the F_1-score, shows a moderately positive correlation between phoneme coverage and segmentation quality at slightly above 0.5.

Overall these results suggest that a higher phoneme coverage should be expected to positively influence the performance of the trained recognizer, although it would be uncertain to what degree.

4.3.2 Oversegmentation

One of the potential metrics evaluated early into the experiments was *oversegmentation*, as defined in [ScWE09]:

$$OS = \left(\frac{\#hypothesizedboundaries}{\#trueboundaries} - 1 \right) * 100$$

As mentioned in section 4.1, oversegmentation turned out to not be useful as a metric due to the behaviour of the function, as well as the complete disregard for the positional accuracy of the hypothesized boundaries. However it is still interesting to observe how various segmenters behave with regard to the number of generated segments, with an oversegmentation value of 0 meaning that a recognizer has hypothesized the exact number of segments present in the ground truth.

Figure 4.10 and table 4.8 show the oversegmentation measurements for decodings with both the multilingual recognizer and the monolingual German one, as well as those for the English baseline system.

We can see that for the cross- and multilingual decoders the optimum lies somewhere between the results for SP settings 0 and 5. For the English baseline system an SP

Language	SP -10	SP -5	SP 0	SP 5	SP 10	SP 20	SP 30
EN	0.3901	0.1435	-0.0107	-0.0950	-0.1548	-0.2484	-0.3258
M5	0.7182	0.3510	0.0916	-0.0651	-0.1615	-0.2943	-0.3914
DE	0.5829	0.2815	0.0631	-0.0587	-0.1372	-0.2458	-0.3293

Table 4.8: Oversegmentation of all segmenters on English Euronews audio

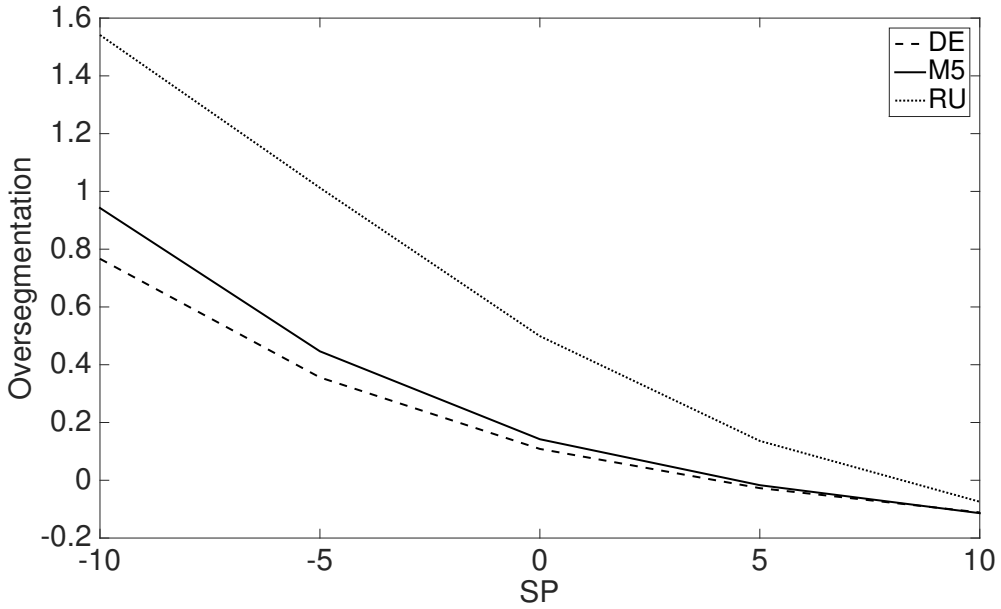

Figure 4.11: Oversegmentation of 3 different segmenters on Basaa audio

value of 0 produces a number of segments that is very close to that indicated by the ground truth, which is to be expected.

Figure 4.11 shows oversegmentation values for our decodings on Basaa. In order to improve readability, the monolingual recognizers that are not Russian are represented solely by German, as they behave rather similarly. Full results can be found in table 4.9.

Language	SP -10	SP -5	SP 0	SP 5	SP 10
M5	0.9427	0.4464	0.1421	-0.0170	-0.1139
DE	0.7667	0.3560	0.1085	-0.0272	-0.1113
FR	0.6714	0.2287	-0.0132	-0.1295	-0.1983
IT	0.7328	0.3281	0.0780	-0.0539	-0.1340
TR	0.8379	0.4284	0.1408	-0.0241	-0.1241
RU	1.5415	1.0125	0.4991	0.1363	-0.0744

Table 4.9: Oversegmentation of all segmenters on Basaa audio

We see that the recognizer trained on Russian data tends to produce considerably more segments than the remaining monolingual recognizers or the multilingual one. The average oversegmentation of monolingual recognizers other than Russian in these experiments on Basaa was 8.05% (for a neutral SP value of 0), whereas for

Russian it is 49.9%. Because of this deviating behaviour, performance results for Russian might not be entirely comparable to scores of other systems. However, it is also possible that the trained model simply leads to additional sub-segmentation of otherwise correct segments, due to some phonetic properties of the Russian language. This could be less problematic for the task of automatically segmenting audio from an unseen language for which no "true" phonetic structure has been previously determined. Further investigation would be required to test this hypothesis.

When calculating the correlation between oversegmentation and performance (see table 4.10) for all systems on Basaa, the Pearson coefficient works out to -0.54 for precision, 0.97 for recall and 0.75 for the F_1-score. This heavy positive correlation between number of segments generated and performance obviously stems from the Russian segmenter with its aberrant behaviour having performed as well as it did. After removing Russian from the calculation, the PCC becomes -0.79, 0.72 and -0.22 (for precision, recall and F_1-score, respectively). This matches with the nature of our metrics: Precision favours fewer segments, Recall more, and the F_1-score as a weighted average of the two should be expected to not show a strong correlation either way.

System	F-Score	Overseg. (%)
M5	**0.5036**	**14.2**
DE	0.4900	10.9
FR	0.5158	-1.3
IT	0.5000	7.8
RU	0.5463	49.9
TR	0.5195	14.8

Table 4.10: Comparison of F-Score and Oversegmentation on Basaa

4.4 Summary and Discussion

In this chapter we have presented and discussed the results of our approach. Precision, Recall and F_1-scores were employed as metrics in order to evaluate the quality of the segmentations generated by our mono- and multilingual systems and we compared them to a gold standard, where possible. Furthermore, we have investigated the effect of a source language's phoneme coverage as well as its tendency to over- or undersegment on the quality of these segmentations.

Initial experiments without language models on English Euronews test data have confirmed the underlying expectation that a multilingually trained system will outperform a monolingual one across language boundaries. F_1-scores of the recognizer trained on 5 different source languages showed a noticeable improvement over its single-language German counterpart, while its perfomance was still significantly lower than that of the baseline system trained on the target language. However, the same expected results could not be observed when applying these segmenters to a different set of audio data of the same language (TIMIT). Here behaviour of metrics across parameter settings seemed erratic, with comparisons among the different systems appearing equally random. Initial theories as to the audio mismatch between

noisy training and clean test data or the phonetic mismatch between accents being responsible could not be corroborated, since further experiments on African audio with at least similar degrees of mismatch did not behave in the same manner.

We also attempted to add additional information in form of a language model to the segmentation process in order to investigate its effect on performance. Unfortunately these models, trained on phoneme sequences found in the training data, rather than word sequences, did not contribute positively to segmentation quality in a significant way. Instead they increased decoding times by a large factor when using the same search settings as before. As a result we must assume that the manner in which these multilingually mixed models were trained is not helpful for the intended task. Whether other approaches that are not based on mixed models would fare better is a question that is left for future work.

Phoneme coverage of a source language on the target was found to have a moderately positive correlation with the quality of the segmentations generated by a system trained on that source language. This result meets intuitive expectations when performing cross-lingual experiments that are purely concerned with phonetic units, not semantic content. The correlation between oversegmentation and performance also confirmed intuition, but only after the Russian recognizer was removed from calculations due to its tendency to heavily oversegment when compared to both, other segmenters at the same settings as well as the ground truth.

Overall the experiments showed that the chosen approach is functional, even though further experiments are required in order to optimize parameters and source language choice, explore different ways of combining source language models, as well as investigating unexpected behaviours encountered while performing the trials.

5. Summary and Outlook

In this work we attempted to utilize multilingually trained speech recognizers in order to cross-lingually segment audio of a previously unseen language on the phonetic level. The approach is meant as a first step to automatically derive information about languages that have previously not been investigated linguistically, and may potentially not have a written representation, e.g. in order to assist the documentation of such languages if they are in danger of going extinct. We used both, a real and a *faux* unseen target language (Basaa and English, respectively). The latter allows for better comparability with existing, regularly trained systems, while the former is closer to the intended scenario of using well-documented source languages to decode a language that is not necessarily closely related.

Several monolingual and multilingually-mixed recognizers were trained using the Janus framework, using context-independent acoustic models. Tests were performed both with and without special language models based on regular n-grams, but applied at the phonetic instead of word level. Examining the performance of the various systems on different target data using precision, recall and F_1-score showed that, generally, a multilingual recognizer performs below a gold standard system trained on the target language, but better and/or more reliably than monolingual systems used in a cross-lingual fashion.

However, the mutlilingual system also exhibited some irregular behaviour on part of the training data that was closer to randomization than to the expected result. The irregularities could not be explained with audio or phonetic mismatch. One of the monolingual recognizers also behaved differently with regard to the number of segments generated, which is likely due to inherent phonetic characteristics of the source language. We also experimented with adding further information in form of special language models trained on phoneme-chains taken from the training data. However this addition did not noticeably contribute to performance, while significantly increasing decoding times.

As for phonetic similarity between source and target languages, correlation between performance and phoneme coverage was weak to moderate. At the very least, as one would expect, phonetic similarity should have *some* positive effect on results, although more tests would be required to ascertain how much.

5.1 Future Work

Apart from advancing to the next steps in the process of documenting a previously unseen language, i.e. clustering the segments found with the approach presented here into a useful phonetic inventory, there are also additional tasks and questions left open due to the scope of this work.

Firstly, using a wider variety of languages would allow to draw conclusions with more certainty. Obviously targeting more languages with the approach would yield more comparable results, especially if a gold standard for these additional languages already exists. Interesting targets include both, other African languages, as well as those from families that are not related to either the Bantu *or* Indo-European groups. Source languages, too, could be more varied, given sufficient available training data. It would also be interesting to see how different combinations of source languages influence performance.

Besides the number and selection of source languages, the manner in which they are combined could also be varied. So far, acoustic models have simply been trained on a mix of training data taken from various languages. Alternatively it may be possible to use monolingual systems to decode target audio individually and then perform a voting between the systems on each frame boundary in order to determine segments. If further experiments show that phonetic similarity does indeed positively correlate with performance, the votes could also be weighted according to phoneme coverage of the individual languages. Determining whether this is the case is another question that could be answered with more diverse pairings of source and target languages.

Another issue that emerged while conducting the experiments presented here was the inconclusive behaviour exhibited by some recognizers on some test data. Using additional corpora that feature different acoustic conditions could allow to determine to what extent such mismatch can negatively impact segmentation quality in our approach. The same holds for the case of some source languages seemingly resulting in systems that produce comparatively large numbers of segments; using other corpora of the same language, or those from phonetically very similar ones might show whether there is a pattern to be found there or not.

Finally, the way phoneme level language models impacted decoding and performance warrants additional attention. While the huge increase in decoding time could be attributed to the large search space resulting from building a lattice on the phoneme instead of the word level, one would expect the additional information to at least noticeably contribute to performance in a positive manner. More tests and possibly also other approaches to training these models both mono- and multilingually could lead to more conclusive results.

Bibliography

[AEEM01] G. Aversano, A. Esposito, A. Esposito und M. Marinaro. A New Text-Independent Method for Phoneme Segmentation. In *44th IEEE 2001 Midwest Symposium on Circuits and Systems*, 2001.

[ASADA+16] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Mayanrd, P. Godard, F. Hamlaoui, D. Idiatov, G. Kouarata, L. F. Lamel, E. Makasso, A. Rialland, M. V. de Velde, F. Yvon und S. Zerbian. Breaking the Unwritten Language Barrier: The BULB Project. In *5th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'16)*, 2016.

[BrFO93] F. Brugnara, D. Falavigna und M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication* 12(4), 1993, S. 357–370.

[BSMB15] P. Baljekar, S. Sitaram, P. K. Muthukumar und A. W. Black. Using Articulatory Features and Inferred Phonological Segments in Zero Resource Speech Processing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[Crys00] D. Crystal. *Language Death*. Cambridge University Press, Cambridge, UK. 2000.

[DaBB52] K. H. Davis, R. Biddulph und S. Balashek. Automatic Recognition of Spoken Digits. *Acoustical Society of America, Journal of* 24(6), 1952, S. 637–642.

[EsAv04] A. Esposito und G. Aversano. A new text-independent method for phoneme segmentation. In *Lecture Notes in Computer Science*, 2004, S. 261–290.

[EsWS07] Y. P. Estevan, V. Wan und O. Scharenborg. Finding Maximum Margin Segments in Speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*, 2007, S. 937–940.

[GAADAB+16] S. S. Gilles Adda, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. V. de Velde, F. Yvon und S. Zerbian. Breaking the Unwritten Language Barrier: The BULB Project. In *Submitted to the 5th International*

Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'16), 2016.

[Garo+93] J. Garofolo und andere. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web download, 1993.

[Gret14] R. Gretter. Euronews: a multilingual benchmark for ASR and LID. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[Jeli09] F. Jelinek. The Dawn of Statistical ASR and MT. *Computational Linguistics* 35(4), 2009, S. 483–494.

[JeMe80] F. Jelinek und R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Workshop on Pattern Recognition in Practice*, 1980, S. 381–397.

[14] R. G. G. Jr. und B. F. G. (Eds.). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, USA. 2005.

[LeRS82] S. E. Levinson, L. R. RAbiner und M. M. Sondhi. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Technical Journal, The* 62(4), 1982, S. 1035–1074.

[LeSF15] P. M. Lewis, G. F. Simons und C. D. Fennig (Hrsg.). *Ethnologue: Languages of the world*. SIL International, Dallas, Texas. 18th. Auflage, 2015.

[Lowe86] B. Lowerre. The HARPY Speech Understanding System. In W. Lea (Hrsg.), *Trends in Speech Recognition*. Speech Science Publications, 1986.

[LWLF+97] A. Lavic, A. Waibel, L. Levin, M. Funke, D. Gates und M. Gavalda. Janus III: Speech-to-Speech Translation in Multiple Languages. In *Acoustics, Speech and Signal Processing (ICASSP), 1997 IEEE International Conference on*, 1997.

[MuBl14] P. K. Muthukumar und A. W. Black. Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, S. 2613–2617.

[NeRo00] D. Nettle und S. Romaine. *Vanishing Voices*. Oxford University Press Inc., New York, NY, USA. 2000.

[QiMi08] Y. Qiao und N. Minematsu. Metric Learning for Unsupervised Phoneme Segmentation. In *Interspeech*, 2008, S. 1060–1063.

[QiSM08] Y. Qiao, N. Shimomura und N. Minematsu. Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons. In *Acoustics, Speech and Signal Processing (ICASSP), 2008 IEEE International Conference on*. IEEE, 2008, S. 3989–3992.

[RLRW79] L. Rabiner, S. E. Levinson, A. E. Rosenberg und J. G. Wilpon. Speaker-Independent Recognition of Isolated Words Using Clustering Techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-27(4), 1979, S. 336–349.

[SaDo62] T. Sakai und S. Doshita. An Automatic Recognition System of Speech Sounds. *Studia phonologica* Band 2, 1962, S. 83–95.

[ScTr03] M. Schröder und J. Trouvain. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology* 6(4), 2003, S. 365–377.

[ScWa01] T. Schultz und A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. *Speech Communication* 35(1-2), August 2001, S. 31–51.

[ScWE09] O. Scharenborg, V. Wan und M. Ernestus. Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *Acoustical Society of America, Journal of* 127(2), 2009, S. 1084–1095.

[SMFW01] H. Soltau, F. Metze, C. Fugen und A. Waibel. A one-pass decoder based on polymorphic linguistic context assignment. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, S. 214–217.

[VTSC+15] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen und E. Dupoux. The zero resource speech challenge 2015. In *Proc. of INTERSPEECH*, 2015.

[WAWBC+94] M. Woszczyna, N. Aoki-Waibel, F. D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Schultz, B. Suhm, M. Tomita und A. Waibel. JANUS 93: Towards Spontaneous Speech Translation. In *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia, 1994.

[WHHS+89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano und K. J. Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech and Signal Processing* 37(3), 1989, S. 328–329.

[WiBe91] I. H. Witten und T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory* 37(4), 1991, S. 1085–1094.

[wKLW07] J. wei Kuo, H. yi Lo und H. min Wang. Improved HMM/SVM methods for automatic phoneme segmentation. In *in Proc. Interspeech*, 2007, S. 2057–2060.

[yLGl12] C. ying Lee und J. Glass. A nonparametric bayesian approach to acoustic model discovery. In *Association for Computational Linguistics, 50th Annual Meeting of the*. Association for Computational Linguistics, 2012, S. 40–49.