Automatic Generation of Pronunciation Dictionaries

For New, Unseen Languages by Voting Among Phoneme Recognizers in Nine Different Languages

Interactive Systems Lab

Carnegie Mellon University, Pittsburgh, PA
Universität Karlsruhe(TH)

Studienarbeit

Sebastian Stüker

Supervisors:
Tanja Schultz
Alex Waibel

April 2002

Abstract

In this report we will describe a data driven approach for creating pronunciation dictionaries for a new unseen target language by voting among phoneme recognizers in nine different languages other than the target language.

In this process recordings of the new language that are transcribed on word level are decoded by the phoneme recognizers. This results in a hypothesis of nine phonemes per time frame, one from every language.

Then two algorithms are described that can map the decoded hypotheses to a pronunciation dictionary entry. These algorithms make use of a confusion matrix based distance measure between the phonemes of the phoneme recognizers and the phonemes of the target language which dictionary is to be created. The confusion matrix is calculated with the help of 500 phonetically transcribed training utterances in the target language.

The phoneme recognizers used in this work were derived from the context independent speech recognizers of the GlobalPhone project.

In order to improve the mapping of the hypotheses of the phoneme recognizers to the dictionary entry we incorporated confidence measures that were derived from word lattices into our algorithms.

Using the proposed algorithms we produced new pronunciation dictionaries for the target languages Swedish and Haitian Creole. The newly created dictionaries were evaluated by comparing the performance of large vocabulary continuous speech recognition systems trained with these dictionaries to reference systems trained with rule based pronunciation dictionaries.

The results of the evaluation show that the process in its current form does not produce pronunciation dictionaries that are accurate enough to train large vocabulary continuous speech recognizers with them. We therefor make suggestions for future work in order to fix the error sources of the process.

Contents

Acknowledgements

I would like to thank Alex Waibel for giving me the opportunity to do the research for this report at the Interactive Systems Lab at the Carnegie Mellon University in Pittsburgh. I would also like to thank Tanja Schultz for arousing my interest in multilingual speech recognition and for her ideas, support, guidance and advice during the course of this project. Furthermore my thanks go to Michael Bett and Rob Malkin for their help with the technical infrastructure in Pittsburgh.

Chapter 1

Introduction

Over the last decade the field of speech recognition has seen enormous progress. Speech recognition can be reliably done on large vocabularies, on continuous speech and speaker independently. The word error rate of these recognizers under certain conditions often is below 10 percent [14]. Large Vocabulary Continuous Speech Recognizers (LVCSR) are commercially available from different vendors. Along with this increased availability comes the demand for recognizers in many different languages that often were not focus of the speech recognition research so far.

It is estimated that today as much as four to six thousand different languages exist [1]. Therefor over the last time increased thought has been given to creating methods for automatizing the design of speech recognition systems for new languages while making use of the knowledge that has been gathered from already studied languages. This new field of research is often referred to as multilingual speech recognition [2].

It is the idea that in that way even for languages with a comparatively small number of speakers LVCSR can be build. These recognizers could then become part of automatic translation systems or speech-driven forms of human computer interfaces.

1.1 Speech Recognition and the JRTk

The experiments in this work were performed with the use of the Janus Speech Recognition Toolkit (JRTk). The JRTk has been developed by the Interactive Systems Laboratories at the Universität Karlsruhe and at Carnegie Mellon University in Pittsburgh [8]. It is part of the JANUS speech-to-speech translation system [7]. A flexible Tcl/Tk script based environment allows building state-of-the-art speech recognizers and provides researchers with a platform that allows them to easily perform new experiments. This toolkit implements an object-oriented approach and unlike other toolkits is not a set of libraries and precompiled modules but a programmable shell with transparent, yet very efficient

objects.

1.2 Pronunciation Dictionaries

One of the core components of a speech recognition system is the pronunciation dictionary. Its purpose is to map the orthographic representation of a word to its pronunciation. In this way it also defines the set of valid phoneme sequences and therefor is a key component in defining the search space of a recognizer. The quality with which it maps the orthography of a word to the way it is pronounced by the speakers has great influence on the performance of the recognition system in two ways. First during the training a false mapping between a word and the phonetic units will contaminate the acoustic models. The models will not describe the actual acoustic that they represent as accurately as if they were only trained with the correct data. Second even when the acoustic models are correctly trained an incorrect mapping will falsify the scoring of a hypothesis by applying the wrong models to the calculation.

Often the creation of a pronunciation dictionary is not a trivial task. It can be created manually by an human expert in the modelled language. But especially with large vocabulary recognizers in which we deal with 60,000 or more words this approach can be very expensive and time consuming and therefor is often not a feasible option. So the process has to be at least in part be automatized. With sufficient knowledge of the target language one can try to build a set of rules that map the orthography of a word to its pronunciation. For some languages this might work very well for others this might be almost impossible. Croatian and Russian are examples for languages with a very close graphem to phonem relation. Thus comparatively few rules suffice to build a pronunciation dictionary containing the canonical pronunciations of the words. So called logoraphic scripts are good examples for languages with no grapheme to phoneme relation. In these scripts a graphem, called a logogram, stands for a single word. A prominent example is the Chinese Hanzi. In Hanzi a logogram does not only stand for a single word but the word it stands for is also dependent on the context. Thus it is virtually impossible to create a dictionary using a rule-based approach. Linguistic experts also often tend to write down the correct, the canonical pronunciation of a word. But this pronunciation may vary greatly from the one that is applied by real speakers. This is especially true for spontaneous speech but applies to other forms such as read speech as well, e.g. due to dialects, accents or the contemporary development of a language.

Sloboda[3] has shown how to improve existing dictionaries for well known languages and how to add new variants whenever needed using a data-driven approach. He used an already existing recognizer with a good performance to find new pronunciation variants by applying the recognizer to the available training data. In this work we will try to build up a dictionary from scratch for a new language that has not been subject to speech recognition research yet. We also assume that we only have a limited amount of training data available for that new language.

1.3 The GlobalPhone Project

This work draws a lot of its resources from the GlobalPhone project [2]. In 1996 Schultz started working on multilingual speech recognition. For her work she needed a database in many different languages that fulfilled the following requirements:

- The languages that are most important for speech recognition according to the number of speakers and their economic or political relevance are covered.
- As many of the phonemes that are used by humans to communicate as possible are covered.
- The speakers of the database are representative for the native speakers of their language. That includes attributes such as gender, age, and level of education.
- The transcribed material is large enough to train robust acoustic models.
- Large additional texts with millions of words are present for calculating a language model.
- The acoustic quality of the material is uniform so that language specific differences can be extracted from the results obtained by the performed experiments.
- All languages are collected with the same type of speech (e.g. spontaneous, read or colloquial speech as a monologue or dialog).
- The data for all languages are similar with respect to their semantics.

At that time no database existed that would have fulfilled all those requirements. Therefor speech and text data for fifteen different languages were collected. The data collection and training has been done in a uniform process to insure the comparability of the resulting recognition systems. The training was also subject to automatization in order to reduce the time needed for building a recognition system. The resulting recognizers were then combined to form a multilingual recognizer that can decode multilingual texts.

With the use of these systems experiments were performed to analyze the possibilities of creating recognizers for new languages using the resources and the knowledge of the already existing systems.

1.4 Objective

In this report we will describe a data-driven approach for creating pronunciation dictionaries that makes use of phoneme recognizers in eight different languages from the GlobalPhone project plus a multilingual phoneme recognizer that makes use of the combined phonemes from seven languages that were

trained with recordings from all seven languages. The creation process will require recordings of the target language as input that are transcribed and segmented on word level but will not require any further linguistic knowledge. In addition we will need a small amount of training data that is segmented on word level and transcribed on phoneme level. This material is needed for calculating a distance measure between the phonemes of the target language — that is the language for which a dictionary is created — and the phonemes of the GlobalPhone phoneme recognizers. In case that no training material that is transcribed on phoneme level is available one could use any kind of distance measure, e.g. one that relies on linguistic knowledge about the phonemes of the target language.

For the purpose of evaluation we also trained speech recognizers for two languages that sofar have not been thoroughly studied with regard to speech recognition. These two languages are Haitian Creole and Swedish. For each language three recognizers were trained. The first recognizer for each language was trained with the use of a rule-based dictionary. These recognizer serves as a baseline for the second and third recognizer that were trained with the use of the dictionaries that have been newly created with the algorithms described in this report.

Chapter 2

Dictionary Creation

In this chapter we will describe the algorithms with which we created four pronunciation dictionaries. The creation process made use of phoneme recognizers in nine different languages.

The two languages for which pronunciation dictionaries were created are Creole and Swedish. Creole was chosen because it shows a close grapheme to phoneme relation. So a relatively small set of rules suffices to create canonical pronunciations. Swedish on the other hand shows a more loose relation of graphemes to phonemes. Thus it is a much more difficult task to create a rule based dictionary. In the following sections we will refer to these languages as the target languages. We further refer to the phoneme set of a target language as $\{T_1, T_2, \ldots, T_v\}$

This chapter gives an outline of the process from a more theoretical and formal point of view. In chapter 3 we will describe the actual experiments that were performed and how the new dictionaries were evaluated. Chapter 3 will also describe the baseline recognizers used for evaluation, the phoneme recognizers used in the following process and will give some background information on Swedish and Creole.

2.1 Outline of the Creation Process

The creation of the pronunciation dictionaries required several steps which are illustrated in figure 2.1 on page 11. In this section we will give an overview of the process while the single steps will be discussed in more detail in the following sections.

The basis for the creation process was formed by nine phoneme recognizers that came from the GlobalPhone project and were slightly improved. The phoneme recognizers were trained for the languages Chinese (CH), Croatian (CR), French (FR), German (GR), Japanese (JA), Portuguese (PO), Spanish (SP) and Turkish (TU). In addition to them a multilingual phoneme recognizer on seven languages was used (MM7). From now on we will refer to these

recognizers either as the phoneme recognizers or simply as the decoders. We will call the languages the GlobalPhone languages. This includes the language mix that is the basis for the MM7 phoneme recognizer. Thus we define the set of GlobalPhone languages $LID = \{CH, CR, FR, GR, JA, PO, SP, TU, MM7\}$. We further define that each of the languages $lid \in LID$ possesses the phoneme set $\{P_{lid_1}, P_{lid_2}, \ldots, P_{lid_{v_{lid}}}\}$.

With the use of these recognizers a confusion matrix was calculated that showed the confusability of the phonemes in the GlobalPhone languages with Creole and Swedish phonemes. This matrix served as a distance measure between the Creole and Swedish phonemes on the one side and the phonemes of the phoneme recognizers on the other side.

The phoneme recognizers were also used to decode audio files of the words whose pronunciations were to be modelled. Thus for every target language word that was subjected to grapheme to phoneme conversion one or more hypotheses in every of the GlobalPhone languages were created.

Then with the help of the confusion matrix the decodings of the target words in the different languages were used to create the pronunciations of the decoded words. As a further addition to improve the decision making process the confidence of the single decoders on the hypotheses that they produced where taken into account when deriving the pronunciations from them.

2.2 The Phoneme Recognizers

The phoneme recognizers were derived from context independent recognizers of the GlobalPhone project that were the result of an intermediate stage in the training.

The MM7 recognizer is somewhat different from the other recognizers. MM7 stands for multilingual mixed in seven languages. Multilingual mixed means that the phoneme sets of seven languages were combined into an universal phoneme set. In order to do so phonemes from different languages that share the same IPA symbol were combined into a language independent phoneme. Now the acoustic models for the phonemes in this global phoneme set were trained with the training material from all the original languages. The resulting recognizer can recognize phonemes from the seven languages without prior knowledge about the language [11].

2.3 Transcriptions on Word Level

For the calculation of the confusion matrix and for the actual creation of the pronunciation dictionary, transcriptions and audio recordings on a word level were needed. However the GlobalPhone project only provided transcriptions and recordings on an utterance level. The Swedish and Creole baseline recognizers provided suitable labels that were used to automatically segment the training data of both recognizers on a word level. The term labels here refers to

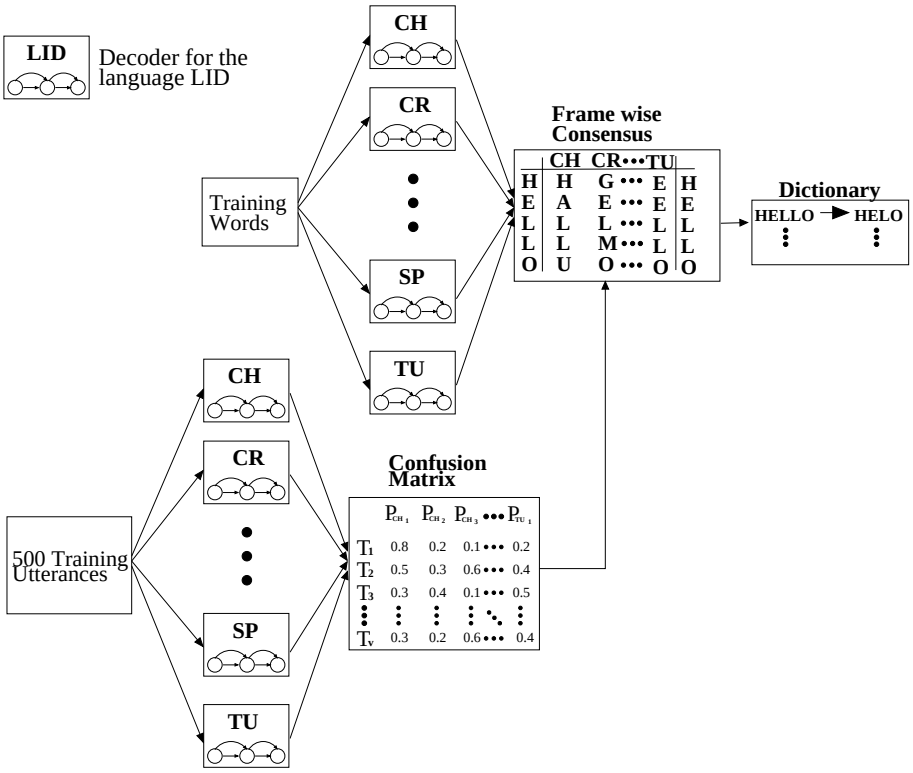

Figure 2.1: The dictionary creation process

the path of a Viterbi alignment as given by the JRTk. The information about the boundaries of the phonemes within the words were taken from the labels as well and were stored in a database together with the words.

2.4 Building a Confusion Matrix

From the segmented training data 500 training utterances from each target language were randomly selected. These words were decoded by each of the phoneme recognizers. The hypotheses plus the information about the hypothesized phoneme boundaries within the words were stored in a file. Using them and the information about the reference words and their phoneme boundaries a phoneme confusion matrix was calculated. In order to do so the references and the hypotheses were compared frame by frame. For every frame it was kept track of how many times a phoneme in one of the target languages had been hypothesized as phoneme P_{lid_i} in one of the GlobalPhone languages. P_{lid_i} means the i^{th} phoneme in language lid where lid is one of the GlobalPhone languages.

	T_1	T_2	$\cdots$	T_v
P_{CH_1}	$1 - \frac{c^{CH}_{1,1}}{norm_{CH,1}}$	$1 - \frac{c^{CH}_{1,2}}{norm_{CH,1}}$	$\cdots$	$1 - \frac{c^{CH}_{1,v}}{norm_{CH,1}}$
P_{CH_2}	$1 - \frac{c^{CH}_{2,1}}{norm_{CH,2}}$	$1 - \frac{c^{CH}_{2,2}}{norm_{CH,2}}$	$\cdots$	$1 - \frac{c^{CH}_{2,v}}{norm_{CH,2}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P_{TU_n}	$1 - \frac{c^{TU}_{n,1}}{norm_{TU,n}}$	$1 - \frac{c^{TU}_{n,2}}{norm_{TU,n}}$	$\cdots$	$1 - \frac{c^{TU}_{n,v}}{norm_{TU,n}}$

Figure 2.2: The confusion matrix

These counts make up a matrix where each entry gives the number of times a reference phoneme had been confused with a GlobalPhone phoneme. Then each entry was normalized by dividing it through the number of occurrences of the corresponding GlobalPhone phoneme. As a final step the normalized confusion counts were subtracted from 1 to transform them into a distance measure. A formal definition of the matrix is given in figure 2.2. Here T_i stands for the i^{th} reference phoneme of the target language. P_{lid_i} refers to the i^{th} phoneme in language lid. $c^{lid}_{i,j}$ is the number of times the phoneme T_j has been recognized as P_{lid_i}, while $norm_{lid,i} = \sum_{j=1}^{v} c^{lid}_{i,j}$ is the normalization factor. As a result we get a Matrix

$$M = \left(1 - \frac{c^{lid}_{i,j}}{norm_{lid,i}}\right)^{j}_{lid,i} = \left(m_{lid_i,j}\right) \tag{2.1}$$

Using this matrix a distance measure between the GlobalPhone phonemes and the target phonemes is defined:

$$dist(P_{lid_i}, T_j) = m_{lid_i,j} \tag{2.2}$$

2.5 Decoding with the Phoneme Recognizers

The phoneme recognizers as briefly mentioned in section 2.2 and described in detail in section 3.2 were also used to decode all words in the training data for the target languages. In addition to the hypotheses and the boundaries of the phonemes the word lattices for every decoded word were saved for later use. So for every occurrence o of every target word w_i different hypotheses $H_{lid}(w^o_i) = P^{lid}_{j_1}(s_1, e_1)\ P^{lid}_{j_2}(s_2, e_2) \ldots P^{lid}_{j_m}(s_m, e_m)$ were decoded, one for every language. Here with $lid \in LID$ and the above definitions of the phoneme sets of the GlobalPhone languages $P^{lid}_{j_k}$ represents the phoneme from the language

	Frames			
lid	1	2	$\cdots$	m
CH	$P_{CH_{i_1}}$	$P_{CH_{i_2}}$	$\cdots$	$P_{CH_{i_m}}$
DE	$P_{DE_{i_1}}$	$P_{DE_{i_2}}$	$\cdots$	$P_{DE_{i_m}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
TU	$P_{TU_{i_1}}$	$P_{TU_{i_2}}$	$\cdots$	$P_{TU_{i_m}}$
Consensus	T_{i_1}	T_{i_2}	$\ldots$	T_{i_m}

Figure 2.3: Finding a Frame Wise Consensus

lid that has been recognized to be the *kth* phoneme in the sequence. It starts at frame s_k and ends at frame e_k.

2.6 Deriving a Pronunciation

For every target word occurrence whose audio recordings were decoded by the phoneme recognizers a pronunciation was derived in a three steps process.

1. **Frame wise consensus:** For every single frame a target phoneme is decided upon that would match the decodings for this frame best. This results in a list of target phonemes, one for every frame of the given word. We call this list a frame wise consensus. In addition a confidence on this frame wise consensus is calculated. This step is explained in detail in subsection 2.6.1.

2. **Pronunciation variants:** In this step every frame wise consensus found in step 1 is transformed into a pronunciation variant. The algorithm used in this step is explained in subsection 2.6.2.

3. **Selecting variants:** After steps one and two for every word in the training set at least one but possibly many different pronunciations are given depending on the number of occurrences of the word in the training recordings. Thus in this step for every word a final dictionary entry is formed from its set of pronunciations that were found so far. This entry can consist either of one pronunciation for the word or can contain multiple pronunciation variants. Refer to subsection 2.6.3 for further explanations.

2.6.1 Finding a Frame Wise Consensus

The task to perform in this first step is to take the decodings as described in section 2.5 and find a phoneme in the target language for every frame. This

problem is illustrated in figure 2.3 on page 13. Since we are now dealing with the decodings on a frame level in this figure $P_{lid_{i_k}}$ refers to the phoneme that has been decoded by language *lid* for the frame k. The target language has the phoneme set $\{T_1, T_2, \ldots, T_v\}$ and so $T(k) = T_{i_k}$ means the phoneme that has been decided upon for the frame i. Given that the decoded word is m frames long the final consensus C for a word occurrence w_i^o is the concatenation of all the consens for the individual frames:

$$C(w_i^o) = \bigoplus_{i=1\ldots m} T(i)$$

This notation will also be used in the following subsections that describe the different ways to calculate such a consensus.

Two different algorithms to solve this problem were implemented. Both take the decodings by the GlobalPhone languages for one word occurrence on frame level as an input and return a sequence of target phonemes — one for every frame of the target word. In addition to the creation of a consensus every algorithm also outputs a measure of confidence for the consensus found.

2.6.1.1 The winner takes it all

This algorithm tries to optimize the distance between the chosen target phoneme and the hypothesized phonemes locally. First every hypothesized phoneme is mapped to the target phoneme with the smallest distance using the distance measure from equation (2.2). This results in a list of target phonemes for every frame. From this list the target phoneme with the most occurrences is chosen for this frame. In case that there are more than one phoneme with a maximum number of occurrences the first phoneme is chosen among them. The average number of occurrences of the maximum phoneme per frame serves as a confidence measure for the consensus found. So let

$$nT(P_{lid_{i_k}}) = T_k \text{ with } dist(P_{lid_{i_k}}, T_k) = \min_{T_j \in \{T_1,\ldots,T_v\}} \{dist(P_{lid_{i_k}}, T_j)\} \quad (2.3)$$

be the function that maps a GlobalPhone phoneme to the nearest target phoneme. We further introduce an indicator function $\delta(T_i, T_j)$

$$\delta(T_i, T_j) := \begin{cases} 1 & : \quad T_i = T_j \\ 0 & : \quad T_i \neq T_j \end{cases} \quad (2.4)$$

with which we now can define a function $numb(T_k, p)$ that gives the number of occurrences of a target phoneme T_k for a given frame p:

$$numb(T_k, p) := \sum_{lid \in LID} \delta(T_k, nT(P_{lid_{i_p}})) \quad (2.5)$$

Now for every frame p the target Phoneme $T(p)$ is chosen that has the most occurrences.

$$T(p) = T_{max} \text{ with } numb(T_{max}, p) = \max_{T_i \in \{T_1,\ldots,T_v\}} \{numb(T_i, p)\} \quad (2.6)$$

Assuming that the hypothesis for the current word occurrence w_i^o is m frames long the confidence for the consensus C is defined as:

$$conf(C(w_i^o)) = \frac{\sum_{i=1}^{m} numb(T(i), i)}{m} \tag{2.7}$$

2.6.1.2 Distance minimization with sliding window

This algorithm tries to optimize the distance between the target and the decoded phonemes on a neighborhood of the frame whose consensus is to be found. So for every frame the target phoneme is chosen that has the smallest cumulated distance to the hypothesized phonemes of the given frame and its neighbors. The size of the neighborhood can be selected and the distances can be weighted for every frame individually. So the cumulated distance for a frame p and a target phoneme T_j is defined as

$$distNeighbor(p, T_j) = \sum_{k=p-l_l}^{p+l_r} (g_{k-p} \sum_{lid \in LID} dist(P_{lid_{i_k}}, T_j)) \tag{2.8}$$

where l_r is the size of the neighborhood in frames to the right and l_l the size of the neighborhood to the left of the current frame that will be included in the calculation of the distance. g_i $(i \in \{-l_l, \ldots, l_r\})$ are the weights for the distances of the frames in the neighborhood. Thus given a word occurrence for every frame p the target phoneme $T(p)$ is chosen that minimizes this distance.

$$T(p) = T_{min} \text{ with } distNeighbor(p, T_{min}) = \min_{T_i \in \{T_1, \ldots, T_v\}} distNeighbor(p, T_i) \tag{2.9}$$

As a confidence measure for the consensi found by this algorithm we take the difference between the maximum possible distance per frame and the average distance of the found target phonemes to the GlobalPhone phonemes per frame. Since the maximum distance in our case is 1.0 the maximum distance per frame is the sum of the weights times the number of GlobalPhone languages times the length of the word occurrence in frames. So the complete formula for the confidence is:

$$conf(C(w_i^o)) = (|LID| * m * \sum_{i=-l_l}^{l_r} g_i) - \frac{\sum_{i=1}^{m} distNeighbor(i, T(i))}{m} \tag{2.10}$$

2.6.2 Transforming the Frame Wise Consensus into a Pronunciation Variant

The frame wise consensus gives a target phoneme for every frame. For the entry in the dictionary however one is only interested in a sequence of single phonemes. So the consensus has to be transformed into a pronunciation variant.

This is done by scanning the consensus from left to right. During the scan silence phonemes are eliminated. Phonemes that do not stay the same for a fixed

$$T_0T_0T_3T_3T_3T_3T_7T_7T_{12}T_{12}T_{12}T_{12}T_6T_6T_0T_0T_0 \Longrightarrow T_3T_{12}$$

$$SilSilSilHHHHEEEEALLLLAOOOOASilSil \Longrightarrow HELO$$

Figure 2.4: Transformation of a consensus into a pronunciation variant

number of frames, e.g. four frames, are eliminated as well. In the remaining consensus the sequences of consecutive equal phonemes are collapsed into a single phoneme.

An example is given in figure 2.4 on page 16. In this example T_0 represents the silence phoneme and the other phonemes have to be at least four frames long in order to be taken over into the pronunciation variant.

2.6.3 Selecting the Pronunciation Variants

For every occurrence of a word in the training set a pronunciation was calculated. Thus the process so far possibly created many different pronunciation variants for every word. From this set of pronunciation variants a final entry in the dictionary had to be created. A first approach to make a selection is to take a fixed number n of variants for every word into the dictionary. This selection can be done in differen ways. Possible algorithms are:

1. Select the variants with the most occurrences. In case of ambiguities (e.g. all variants have the same number of occurrences) choose among the ambiguous variants the ones with the higher average confidence.

2. Select the variants with the highest confidences. In case of ambiguities (e.g. all variants have the same confidence) select among the ambiguous variants the ones with the higher number of occurrences.

3. Weight the number of occurrences of a variant with the averaged confidence and take the best. In case of ambiguities choose among the ambiguous variants the ones with the higher confidence.

Here one sees the number of occurrences of a variant as a measure of confidence for this variant. However selecting a fixed number of variants for every word neglects the fact that for words with many occurrences it would be desirable to select more pronunciation variants since it is likely that more pronunciation variations have been seen in training. So we modified this approach by taking not a fixed number of pronunciations but selecting a certain percentage of pronunciation variants of the distinct variants that have been produced by the consensus process. However in order to avoid too extensive entries we limited the number of variants by a constant N_{max}. Of the listed selection criteria we chose the third one feeling that it provides a good mixture of using both sources of knowledge: the number of occurrences of a variant and the confidence associated with it.

Instead of taking a fixed number or certain percentage of pronunciation variants for every word one could also pick the variants by applying a threshold to either the number of occurrences or the confidence of the single variants. By selecting all the variants above the threshold one sets a certain minimum standard for the quality of the pronunciation. If the standard is not met one would rather not model a pronunciation at all than to create a bad one. However since we only have a limited training material with only a comparatively small number of words deciding not to model a pronunciation at all would increase the out of vocabulary rate even further.

2.7 Incorporation of Confidence Measures

The introduced process heavily relies upon the output of the phoneme recognizers. But as we will see in chapter 3 the performance of our phoneme recognizers is far from being perfect. Some of the phoneme recognizers do not even reach a phoneme accuracy of 60%. By using seven phoneme recognizers we try to counteract this effect hoping that the mistakes made by few recognizers will be overlapped by the results of the others. But still one can assume that the results of the algorithms that find a frame wise consensus are largely influenced by the performance of the phoneme recognizers and that a lower performance of the recognizers can produce undesirable results. Traditional speech recognition faces the same problem. LVCSR systems are far from being perfect and it is often desirable to know when a mistake in the recognition process has been made. E.g. this might be helpful in a dialog system where the system can try to verify an input by asking the speaker for further information or for repetition of the input. One way to predict or detect errors in hypotheses produced by recognition systems is to apply a measures-of-confidence to the hypotheses. A low confidence on a word in a hypothesis would then be a good indication that a mistake has been made. For our task one cannot speak of mistakes made during the decoding process since it is almost impossible to say at the point of the decoding what the desirable output would be. However we tried to apply the same confidence measures used for detecting errors in hypotheses to the frame wise consensus process. The idea is that the confidence measures can be used to control the influence that the hypothesis of a language has on the arbitration process. A high confidence would imply a greater influence, a lower confidence a smaller influence.

2.7.1 Different Kinds of Confidence Measures

The Janus Speech Recognition Toolkit uses word lattices as a representation of a set of alternative hypotheses. A word lattice is a directed graph which vertices represent words and which links induce a possible succession of words. Thus every path through the lattice represents an alternative hypothesis. Kemp and Schaaf have proposed and examined ways in which confidence measures can be derived from word lattices [4] [5]. Two features that can be derived

from a lattice were selected and incorporated into the decision making process. The first confidence measure is called Gamma. To calculate this feature the lattice is interpreted as an HMM. The nodes of the lattice correspond to the states of the HMM while the links correspond to transitions. The emission probability for the states are the acoustic scores of the corresponding words and the transition probabilities are given by the language model. Since we do not use a language model for the phoneme recognizers the probabilities for all transitions are equally distributed. With this interpretation a forward-backward algorithm can be computed over the word lattice which assigns an a posterior probability to each of its nodes and links. Experiments by Kemp and Schaaf have shown that Gamma performs very well.

Another confidence measure or actually a whole set of confidence measures can be derived from the lattice by counting the number of alternative words that are allowed per unit of time respectively per frame. For LVCSR the search space is so huge that unlikely hypotheses have to be pruned away. In a time segment were the probability of a word W_i is considerably higher than the probability of the other words most of those words will be pruned away. Since the lattice is a representation of an already pruned search space the number of alternative words per time segment should be low. If a small number of words with a relative high probability does not exist, but rather many words have a similar probability no effective pruning can take place. A large number of alternative hypotheses with similar probabilities implies a higher probability of error. Therefor one should expect that a high number alternative words per time frame in a lattice implies a lower confidence. Schaaf and Kemp have shown in [4] that features derived from the hypotheses density in word lattices can be used to build a classifier for confidence tagging. From the features that were proposed in [4] TAve was chosen. TAve of a word w is the average number of alternative links per time frame over its time span.

Since during the process of decoding the target language the word lattices were saved these confidence measures were later calculated and the decodings were then tagged with them. We will now discuss how the confidences were incorporated into the process of finding a frame wise consensus. For this we will assume that a confidence measure $conf(P, p)$ is given that gives the confidence of the phoneme P that has been hypothesized for the frame p.

2.7.1.1 The Winner Takes it All

The algorithm "The Winner Takes it All" as described in subsection 2.6.1.1 can produce an ambiguous result. When selecting the phoneme with the most occurrences after mapping the hypothesized phonemes to their nearest target language phoneme it can be the case that more than one phoneme can have a maximum number of occurrences. Among this set of candidate phonemes one has to make a selection. So far we chose the first phoneme among this set. Now we try to resolve this ambiguities by making use of the confidence assigned to the hypothesized phonemes. Therefor we transfer the confidence of the hypothesized phoneme to the target language phoneme that it is mapped

to. Then we can chose among the set of phonemes with a maximum number of occurrences by taking the one with the highest averaged confidence. The confidence of the frame wise consensus can still be calculated as in equation (2.7).

This way of finding a consensus still heavily relies on the number of occurrences of a phoneme. A different approach to incorporate confidence measures would be to change the selection of the target phoneme after the hypothesized phonemes have been mapped to target phonemes. Instead of maximizing the number of occurrences one selects the phoneme that maximizes the sum of the confidences over the selected phoneme. So using equations (2.3) and (2.4) we define the function $sumConf(T_k, p)$ that gives the accumulated confidence of the GlobalPhone phonemes that have been mapped to target phoneme T_k for frame p:

$$sumConf(T_k, p) := \sum_{lid \in LID} conf(P_{lid_{i_p}}, p) * \delta(T_k, nT(P_{lid_{i_p}})) \tag{2.11}$$

In analogy to equation (2.6) we now choose the target phoneme that maximizes the accumulated confidence:

$$T(p) = T_{max} \text{ with } sumConf(T_{max}, p) = \max_{T_i \in \{T_1, \ldots, T_v\}} \{sumConf(T_i, p)\} \tag{2.12}$$

In analogy to equation (2.7) we now calculate the confidence of the frame wise consensus as the average cumulated confidence:

$$conf(C(w_i^o)) = \frac{\sum_{i=1}^{m} sumConf(T(i), i)}{m} \tag{2.13}$$

2.7.1.2 Sliding Window

To incorporate the confidence measures into the algorithm described in 2.6.1.2 we weighted the distances between a hypothesized phoneme and a target phoneme with the confidence of the hypothesized phoneme. Here the idea is that we want to decrease the influence of the hypothesized phoneme in case of a low confidence. Since we are minimizing the sum of the distances between the hypothesis phonemes and the target phoneme we can do this by reducing the distances from a phoneme that has been hypothesized for this frame to all target phonemes. Thus the the distances from hypothesized phonemes with a higher confidence will add more to the sum of the distances and thus will have more influence on the decision for a target phoneme. We modify equation (2.8) to:

$$distNeighbor(p, T_j) = \sum_{k=p-l_l}^{p+l_r} (g_{k-p} \sum_{lid \in LID} conf(P_{lid_{i_k}}, p) * dist(P_{lid_{i_k}}, T_j)) \tag{2.14}$$

The rest of the algorithm remains as described in 2.6.1.2. The definition of the confidence stays the same as in equation (2.10). We just use the newly defined

function $distNeighbor$ from equation (2.14) instead of the old definition from equation (2.8).

Let us take a look at an example. We assume we have a target language with a very small phoneme set of only two phonemes $\{T_1, T_2\}$. We further assume that we used two decoders to decode this target language and we now want to find a consensus for the frame p. The first decoder hypothesized the phoneme P_1 for this frame and the second decoder hypothesized the phoneme P_2. A distance measure is given as well. The following matrix shows the distances between the hypothesized phonemes and the target phonemes:

	T_1	T_2
P_1	0.5	0.8
P_2	0.5	0.4

The confidences with which P_1 and P_2 have been hypothesized are:

$$conf(P_1, p) = 0.2 \qquad conf(P_2, p) = 0.9$$

For the sake of simplicity we choose the smallest possible neighborhood for calculating a consensus which is the frame to find a consensus for itself and weight it with 1.0. So using the notation from equation (2.8):

$$l_l = l_r = 0 \quad \text{and} \quad w_0 = 1.0$$

Without using the confidences we would find with the algorithm described in subsection 2.6.1.2 and equation (2.8) that:

$$\begin{aligned}
& distNeighbor(p, T_1) \\
= \;& g_0 * (dist(P_1, T_1) + dist(P_2, T_1)) \\
= \;& 1.0 * (0.5 + 0.5) \\
= \;& 1.0 \\
< \;& 1.2 \\
= \;& 1.0 * (0.8 + 0.4) \\
= \;& g_0 * (dist(P_1, T_2) + dist(P_2, T_2)) \\
= \;& distNeighbor(p, T_2)
\end{aligned}$$

Since the cumulated distance to target phoneme T_1 is smaller than the one to target phoneme T_2 we would take T_1 as the consensus for this frame. Let us see what happens when we incorporate the confidences as described in equation (2.14). Now we end up with:

$$\begin{aligned}
& distNeighbor(p, T_1) \\
= \;& g_0 * (conf(P1, p) * dist(P_1, T_1) + conf(P2, p) * dist(P_2, T_1)) \\
= \;& 1.0 * (0.2 * 0.5 + 0.9 * 0.5) \\
= \;& 0.55 \\
> \;& 0.52 \\
= \;& 1.0 * (0.2 * 0.8 + 0.9 * 0.4) \\
= \;& g_0 * (conf(P_1, p) * dist(P_1, T_2) + conf(P_2, p) * dist(P_2, T_2)) \\
= \;& distNeighbor(p, T_2)
\end{aligned}$$

So the cumulated distance to target phoneme T_1 is larger than the one to T_2. Thus this time we decide upon T_2 as the consensus for this frame.

When not using the confidence we decide upon T_1 because the distance between P_1 and T_2 is so big that it outweighs the fact that P_2 is closer to T_2 than to T_1. However when we use the confidence the large distance between P_1 and T_2 becomes less relevant because P_1 has been hypothesized with such a low confidence. Instead the fact that P_2 is closer to T_2 than to T_1 now dominates the consensus because of the high confidence of P_2.

Chapter 3

Experimental Results

As mentioned before pronunciation dictionaries for the languages Creole and Swedish were produced using the algorithms described in the previous chapter.

In this chapter I will describe the experiments with which these dictionaries were produced and how they were evaluated by comparing the performance of two LVCSRs trained with those dictionaries against the performance of LVCSRs that were trained with the help of rule based dictionaries.

3.1 The Baseline Recognizers

In order to evaluate the newly created dictionaries two LVCSRs were trained with the help of rule based pronunciation dictionaries. Their performance acts as a baseline against which the performance of the new recognizers in the same language can be compared to when being trained with the new dictionaries.

3.1.1 The Creole Recognizer

When we talk about Creole in this work we actually refer to the Haitian Creole that is spoken on Haiti. Estimates of the number of speakers that one can find range from "more than five million" [1] to eight million [10]. It is the Creole language with the most speakers that is derived from French. The official language on Haiti is French. But only a small amount of the population actually uses it in every day life. Instead the great majority of the Haitian citizens uses Creole on a daily basis.

3.1.1.1 The database

The data collection took place on Haiti and was done by a native speaker. The recordings were done with a laptop and a headset with a close speaking microphone. The collected data were read newspaper articles taken from the internet site of the Haiti Progres *www.haiti-progres.com*. The sampling rate

	Rec Length	N_{spk}	N_{utt}	N_w	N_v
Training	4.2 h	15	1762	33223	2986
Cross	0.48 h	8	158	3558	1005
Evaluation	0.58 h	15	211	4488	741
Total	5.3 h	38	2131	41269	4732

Table 3.1: The Creole Data Basis

of the recordings is 16 kHz while using 16 bit per sample. For training and evaluation purposes the data were divided into three sets:

- A training set with 4.2 hours of recorded speech for training the acoustic models
- A cross validation set with roughly half an hour of recorded speech for adjusting the weight of the language model and the word penalty
- An evaluation set with little more than half an hour of recorded speech for evaluating the performance of the final system

Table 3.1 on page 23 gives an overview of the collected data and their separation into the three sets. N_{spk} stands for the Number of speakers, N_{utt} for the number of utterances, N_w for the number of words and N_v for the size of the vocabulary.

Since the preparation time for the data collection was very short only an insufficient amount of text had been downloaded and prepared for recording. This resulted in the fact that almost all texts where read at least twice by different speakers. However for the acoustic training the set of speakers for the three data sets had to be disjunct. The cross validation set, the evaluation set and the texts used for calculating the language model have to be disjunct as well in order to get valid evaluation results. In order to meet this restrictions some recordings could not be used for the construction of the system and had to be excluded from the three sets.

In order to calculate a statistical language model Creole texts were downloaded from the Internet in addition to the texts from the training set. Articles from *www.haiti-progres.com* that were not used in the data collection as well as news articles from other Internet sources were collected. We also found a Creole translation of the New Testament.

From the collected texts two text corpora were created. The first text corpus included all the collected texts while the second corpus excluded the New Testament. Table 3.2 on page 24 gives an overview of the two text corpora. In this table N_{dw} refers to the number of distinct words.

3.1.1.2 The Dictionary

The dictionary was created from the text corpus that excluded the New Testament because of the assumption that the domain of the newspaper articles and

	N_{utt}	N_w	N_{dw}
with Bible	75340	1165476	18338
without Bible	11518	230263	12179

Table 3.2: The Creole Text Corpora

the Bible would vary too much.

Creole shows a very close grapheme to phoneme relation so that approximately 50 rules are sufficient for converting words that do not contain digits. Another 37 rules were used to generate the pronunciation of numbers up to ten thousand. Since the dictionary was created from the text corpus without the Bible it includes 12179 words.

3.1.1.3 The Language Model

The language model was generated on the text corpus that includes the New Testament using the language model tool written by Klaus Ries [9]. The current standard model that this tool produces features trigrams, Kneser/Ney backoff, non-linear interpolation and absolute discounting with separate estimates for entries occurring once or twice. With the default settings and the above described dictionary as a vocabulary the perplexity of the evaluation set was 252.7. The dictionary showed an out of vocabulary (OOV) rate of 3.761%. Of the trigrams 34.37% were used, 38.88% of the bigrams and 26.75% of the unigrams.

3.1.1.4 The Acoustic Training

The acoustic training of the Creole recognizer followed roughly the same steps as the training of the recognizers for the GlobalPhone languages.

The preprocessing is done in the same way as for the GlobalPhone recognizers and consists of a combination of mel scaled cepstrum coefficients and some dynamic features such as the power of the signal and an approximation of the first and second derivative of the mel coefficients. Using an LDA transformation the dimension of the feature vector is reduced to 32. The power spectrum is also subject to vocal tract length normalization.

A Creole speech recognizer that had been developed during an ISL speech lab at the Universität Karlsruhe provided suitable labels. Using these labels we could skip the initialization of the acoustic models and the step of writing initial labels. Instead we trained a context independent fully continuous recognizer right away. From there we made the transition to a context dependent recognizer that used triphones. After collecting and training the triphones they were clustered into a distribution tree and pruned. This step resulted in a semi continuous phonetically tied recognition system with 3000 codebooks [12].

	Rec Length	N_{spk}	N_{utt}	N_w	N_{dw}
Training	17.4 h	79	9406	144700	24165
Cross	2.1 h	9	1207	18455	5746
Evaluation	2.2 h	10	1203	17870	5429
Total	21.7 h	98	11816	181025	35340

Table 3.3: The Swedish Database

3.1.1.5 The Evaluation

The trained recognizer was evaluated using the speakers from the evaluation set mentioned above. The word accuracy is 55.2 percent. 32 percent of the errors were substitutions 7.2 percent were deletions and 5.6 percent were insertions. The high number of substitutions suggests that the acoustic models are far from being optimal. One reason for this can be found in the small amount of available training material.

3.1.2 The Swedish Recognizer

The Swedish recognizer that is used as a baseline was trained by Schultz in the course of the GlobalPhone project [2].

3.1.2.1 The Database

The speech data for this recognizer were collected as part of the GlobalPhone project under the GlobalPhone rules. For the collection newspaper articles from the Internet site of the newspaper Göteborgs-Posten *www.gp.se* were read. The recordings were performed with the use of the portable DAT-recorder TDC-8 from Sony and the Sennheiser HD-440-6 microphone. Later the recordings were transferred onto a computer and resampled to 16 kHz and 16 bit per sample[2]. Table 3.3 on page 25 gives an overview of the collected data.

3.1.2.2 The Dictionary

The dictionary for this recognition systems was created by a combination of a data-driven and a rule-based algorithm. First it is checked whether for a new word its pronunciation or the pronunciation of parts of it is already known. Then the pronunciation of parts that are not known is modelled by a set of 250 rules.

3.1.2.3 The Language Model

The additional texts for the language model were all downloaded from the same site as the texts for the data collection.

Language	Accuracy
Chinese	51.2 %
Croatian	58.9 %
French	53.3 %
German	53.9 %
Japanese	67.4 %
Portuguese	55.0 %
Spanish	67.0 %
Turkish	57.2 %

Table 3.4: Phoneme accuracy of the GlobalPhone phoneme recognizers

3.1.2.4 Acoustic Training

The preprocessing is the same as described for the Creole recognizer. The training itself followed the same steps as the training for the other GlobalPhone recognizers [2].

3.1.2.5 Evaluation

The evaluation of the Swedish recognizer showed a word accuracy of 33.2 percent for the context independent recognizer and an accuracy of 47.3 percent for the context dependent recognizer.

3.2 The GlobalPhone Phoneme Recognizers

The phoneme recognizers were derived from the context independent recognizers of the GlobalPhone project. The context independent recognizers originally were trained with 32 gaussians in the same manner as described in 3.1.1.4. The phoneme recognizers were created by extending the number of Gaussians to 128. Then four iterations of a Viterbi training were performed using the best available labels. In order to turn the resulting recognizer into a phoneme recognizer a dictionary containing only the phonemes of the corresponding language were created as well as a language model that only contained the equally distributed phonemes as unigrams. The new recognizers were evaluated on their corresponding language. Table 3.4 shows the resulting phoneme accuracy for every language.

3.2.1 Decoding

The GlobalPhone phoneme recognizers were then used to decode the acoustic material that was available for Creole and Swedish. The decodings also included the boundaries of the hypothesized phonemes.

3.2.2 Incorporating the Confidence Measures

In order to be able to experiment with different types of confidence measures we saved the word lattices of the decodings. Later the lattices were used to extract the confidence measures described in chapter 2. We further modified the confidence measures to be able to make better use of them for our specific task.

The confidence measure Gamma produces confidences that numerically varies between the languages so that two phonemes that are recognized with roughly the same confidence in either language are possibly assigned completely different Gamma values. We therefor normalized the maximum Gamma value in every language to 1.0. We will call this confidence measure "GammaNorm".

$$GammaNorm(P, lid) = \frac{Gamma(P, lid)}{\max\{Gamma(lid)\}} \tag{3.1}$$

TAve has the property that the higher its numerical value the lower the confidence on the recognized phoneme. However for our purposes it would be useful that a confidence measure with a high value indicates a high confidence and a low value a low confidence. We therefor had to transform the values that we got from TAve. We did this conversion by means of a linear transformation as described in the equation below. We will call the result"TAveLin"

$$TAveLin(P) = \frac{C - TAve(P)}{C} \quad with\ C \gg 0 \tag{3.2}$$

Just like Gamma the numerical value of TAveLin can vary for phonemes recognized with different recognizers though their confidence should be very similar. Again we tried to correct this problem by means of normalization this time assuming that the TAveLin values show an offset that is characteristic for their language. Therefor we subtracted the mean TAveLin value for every language and called the result TAveOffset.

$$TAveOffset(P, lid) = TAveLin(P, lid) - \mu(lid) \tag{3.3}$$

3.3 Construction of the Dictionaries

With the decodings from the phoneme recognizers and the confidence measures derived from the lattices at hand the next step was to calculate the new dictionaries.

3.3.1 The Confusion Matrix

In order to construct the two confusion matrices as described in the previous chapter we took 500 utterances from the decodings for both Creole and Swedish. With these decodings we constructed a confusion matrix for both languages as described in 2.4. These matrices then acted as a distance measure between the GlobalPhone phonemes and the Swedish and Creole phonemes.

3.3.2 Selecting the dictionaries for evaluation

In chapter 2 we proposed the two different algorithms "The Winner Takes it All" and "Sliding Window". Both algorithms can be combined with the above confidence measures as described in section 2.7. In addition to that "Sliding Window" can be configured with the weights for the neighborhood. Since the evaluation of a dictionary requires the training and evaluation of a complete recognizer it is not a feasible option to completely evaluate a large set of dictionaries in order to find the best combination of algorithm, parameters and confidence measure.

We therefor created a larger set of dictionaries that contained a small amount of words and then manually selected among them the four most promising combinations of algorithm and confidence measure.

Due to the lack for a better method of parameter adaption we decided to configure the "Sliding Window" algorithm with two sets of weights:

- $g_{-1} = \frac{1}{2}$, $g_0 = 1$, $g_1 = \frac{1}{2}$
- $g_{-2} = \frac{1}{4}$, $g_{-1} = \frac{1}{2}$, $g_0 = 1$, $g_1 = \frac{1}{2}$, $g_2 = \frac{1}{4}$

From now on we will refer to the algorithm "Sliding Window" configured with the first set of parameters as "Sliding-1" and to the algorithm configured with the second set as "Sliding-2". For the algorithm "The Winner Takes it All" we introduced two different ways to incorporate the confidence measures in section 2.7.1.1. We will call the first method in which we resolve the ambiguities using the confidence measure "Winner-1". The second method that was introduced we will call "Winner-2".

"Sliding-1", "Sliding-2", "Winner-1" and "Winner-2" were then combined with the confidence measures GammaNorm, TAveLin, TAveOffset and applied to a reduced set of 18 distinct words with a total of 157 occurences. The manual review of the dictionaries showed that the following combination of dictionaries seemed to be most promising:

Swedish:

- Winner-1 + TAveOffset
- Winner-2 + TAve

Creole:

- Sliding-1 + GammaNorm
- Sliding-2 + GammaNorm

3.3.3 Selecting the pronunciation variants

In section 2.6.3 we described three ways how to select the final pronunciation variants that are produced by the above algorithms. For our evaluation we chose

Dictionary	Context Independent	Context Dependent
Baseline	33.2%	47.3%
Winner-1 + TAveOffset	17.5%	8.9%
Winner-2 + TAve	13.2%	7.7%

Table 3.5: Results of the Swedish dictionaries

Dictionary	Context Independent	Context Dependent
Baseline	45.5%	55.2%
Sliding-1 + GammaNorm	4.7%	4.2%
Sliding-2 + GammaNorm	3.9%	3.4%

Table 3.6: Results of the Creole dictionaries

to take the third introduced method. To limit the number of variants in the dictionary we decided to take the 70 percent best variants for a word but not more than a total of 20 variants per word.

3.4 Evaluation of the new Dictionaries

In order to evaluate the selected dictionaries we trained new recognizers from scratch using the same steps as we did for the prior training with the rule based dictionaries. The performance of the new context dependent and context independent recognizers in comparison to the recognizers trained with the rule based dictionaries for Swedish are shown in table 3.5. The results for Creole can be found in table 3.6.

The results show that the process in its current form is clearly not fit for training speech recognizers. The results of the Creole dictionaries are especially discouraging.

3.5 Analysis

In order to determine the reasons for the poor performance of the newly created dictionaries a detailed error analysis is necessary. However the time frame of this work did not allow such an analysis.

But from the experiences we made while performing the experiments we can make some reasonable assumptions about the sources of errors.

3.5.1 Performance of the Phoneme Recognizers

If we assume that our phoneme recognizers produce a perfect output it can be easily seen that our algorithms for finding the frame wise consensus produce the correct result. A perfect output would be if the GlobalPhone phoneme

recognizers would always hypothesize the phonemes in their language that are closest to the target target language phoneme that was spoken in the recordings.

It is therefor reasonable to assume that one of the main problems is an insufficient performance of our phoneme recognizers.

So it the author's opinion that one focus of future work should be on the improvement of the phoneme recognizers. One way to accomplish this could be to use the training material for the confusion matrix to adapt the recognizers to the target language.

As we described in chapter 2 we tried to use confidence measures to compensate for the poor phoneme accuracy of the recognizers. However after performing the experiments we doubt that using the confidence measures had the desired effect. One would have to verify that it is possible to use the confidence measures the way we do.

3.5.2 Selection of Pronunciation Variants

Another likely source of errors is the selection of the pronunciation variants. As described in this work our process is likely to produce multiple different pronunciations for each word, depending on how often the word occurs in the recordings.

It is especially difficult to select the correct number of pronunciation variants. Selecting too many variants is known to confuse the recognizer. On the other hand if the number of selected variants is too low one is in danger of leaving out a relevant variant.

It probably is necessary to examine methods of accomplishing this task that are different from the one we used.

Chapter 4

Summary

4.1 Summary

In this work we tried to develop an automatic process to produce pronunciation dictionaries for new unseen languages. We made use of phoneme recognizers for already studied languages that came from the GlobalPhone project. Our idea was that by skillfully combining the output of the recognizers on the new language the pronunciation of the new language can be modelled. Two algorithms for this combination were proposed and an attempt was made to enhance them with confidence measures that have already been proven to be effective for existing speech recognizers.

Pronunciation dictionaries for Swedish and Creole were produced using the proposed algorithms and confidence measures. We selected among the resulting dictionaries the most promising ones and evaluated them by comparing the performance of a LVCSR trained with them to the perfromance of two LVCSRs trained with rule based dictionaries.

4.2 Future Work

The results showed that the process in its current form does not yet produce satisfactory results. In order to find the weak points of the algorithms a thorough examination of the sources of errors will have to be done.

However it seems already clear that further work will have to be put into improving the performance of the phoneme recognizers on the unseen languages. One possibly way to do this could be to adapt the phoneme recognizers to the target language using the 500 utterances that right now are needed for calculating the confusion matrix.

List of Figures

List of Tables

Bibliography

[1] Hans Joachim Störig. Abenteuer Sprache. München 1997

[2] Tanja Schultz. Multilinguale Spracherkennung. Karlsruhe, 2000

[3] Tilo Sloboda. Dictionary Learning: Performance Through Consistency. In Proceedings of the ICASSP 1995, Detroit, USA, volume1, pp. 453-456

[4] Thomas Kemp, Thomas Schaaf. Estimating Confidence Using Word Lattices. In Proceedings of the Eurospeech 1997, Rhodes, Greece

[5] Thomas Schaaf, Thomas Kemp. Confidence Measures For Spontaneous Speech Recognition. In Proceedings of the ICASSP 1997, vol. 2, pp. 875ff, Munich, April 1997

[6] F. Metze, T.Kemp, T.Schaaf, T.Schultz and H. Soltau. Confidence Measure Based Language Identification. In ICASSP 2000, Istanbul, Turkey, 2000

[7] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld Puming Zhan. JANUS-III: Speech-To-Speech Translation in Multiple Languages. In Proceedings of the ICASSP 1997, Munich, 1997

[8] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, Matrin Westphal. The Karlsruhe-VERBMOBIL Speech Recognition Engine. In Proceedings of the ICASSP 1997, Munich, 1997

[9] Klaus Ries, Bernhard Suhm, Petra Geutner. http://www.is.cs.cmu.edu/local/janus-lm.doku/janus-lm.doku.html

[10] http://www.haiti.com

[11] T. Schultz and A. Waibel. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. In Proceedings of the ICSLP 1998, vol. 5, pp. 1819-1822, Sydney, November 1998

[12] M.Finke. I.Rogina. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1743-1746, Munich, April 1997

[13] T. Schultz, A. Waibel. Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages. In Workshop on Speech and Communication (SPECOM-1998), pp 207-210, St. Petersburg, Russia, October 1998

[14] David S. Pallet, Jonathan G. Fiscus, John S. Garaofolo, Alvin Martin, and Mark Przybocki. 1998 Broadcast News Benchmark Test Results. In Proceedings of the DARPA Broadcast News Workshop, Herndon, Virginia, February 28-March 3, 1999