MULTIGRAMS FOR LANGUAGE IDENTIFICATION

Stefan Harbeck and Uwe Ohler
Chair for Pattern Recognition
University of Erlangen-Nuremberg
Martensstrasse 3
D-91058 Erlangen, Germany
e-mail: snharbec@informatik.uni-erlangen.de
www: *www5.informatik.uni-erlangen.de/Persons/MA/hb*

ABSTRACT

In our paper we present two new approaches for language identification. Both of them are based on the use of so-called multigrams, an information theoretic based observation representation. In the first approach we use multigram models for phonotactic modeling of phoneme or codebook sequences. The multigram model can be used to segment the new observation into larger units (e.g. something like words) and calculates a probability for the best segmentation. In the second approach we build a fenon recognizer using the segments of the best segmentation of the training material as "words" inside the recognition vocabulary. On the OGI test corpus and on the NIST'95 evaluation corpus we got significant improvements with this second approach in comparison to the unsupervised codebook approach when discriminating between English and German utterances.

1. INTRODUCTION

Language identification has been a field of interest for the last ten years. A wide spread method for language identification is based on the evaluation of phonotactic knowledge which is usually done by using stochastic language models [see Zissman, 1996]. The stochastic language models are trained and evaluated on phoneme sequences, which are extracted out of the speech signals using a phoneme recognizer. This phoneme recognizer has to be trained on transcribed material, which should be as close as possible to the application domain. Recent algorithms use several language specific phoneme recognizers, so the requirements on the training material is much more higher. The development of applications on different databases especially with different signal quality demands for a new transcribed database which is not available for every language.

In contrast we had focused on methods for language identification which require less information about the training material [Harbeck et al., 1997]: We need only a set of signals for each language and no additional transcription. Application on new domains and different signal quality is possible just by recording the samples within this domain and use them to train the new language identification module.

Most of the algorithms we investigated so far are based on the extraction of codebook sequences, which can be trained without any transcription of the training material. These algorithms were much better where discriminating between 13 languages of a military database than a phoneme recognizer which was trained on the OGI corpus. But in recent experiments carried out on a part of the OGI corpus, the phoneme approach was superior to the unsupervised vector codebook approach. The problem of the codebook approach is that it does not model phonotactic knowledge directly because the units extracted do represent only parts of phonemes. Due to the smaller length of our codebook classes with respect to the phonemes, the recognition process was also very error prone, which lead to erroneous phonotactic models. So we searched for acoustic units which are similar to phonemes in an unsupervised way. One method is to search for acoustic homogenous regions e.g. by means of temporal decomposition [Bimbot and Atal, 1991]. Another method presented in this paper is based on information theory.

In the following paper two different approaches are described, which are both based on information theoretic units called multigrams. In the first approach the standard stochastic language model is replaced by the multigram model, in the second the acoustic units which will be used inside the recognizer will be replaced by the multigrams units.

The paper is organized as follows: In the next section an introduction to multigrams is presented. An overview about the base line system based on codebook sequences is given in section 3. The description of the two new approaches based on multigrams follows. In section 5 experiments on a part of the OGI corpus are presented. A conclusion will be given in section 6.

2. MULTIGRAMS

The problem of phonotactic modeling can be interpreted as finding a grammar which fits the training data. As there might exist many consistent grammars, Chomsky wrote in [Chomsky, 1955]:

> In applying this theory to actual linguistic material, we must construct a grammar of the proper form... Among all grammars meeting this condition, we select the simplest. The measure of simplicity must be defined in such a way that we will be able to

[1]This work was supported by MEDAV GmbH.

> evaluate directly the simplicity of any proposed grammar... It is tempting, then, to consider the possibility of devising a notational system which converts consideration of simplicity into consideration of length.

Chomsky's idea about a relationship between quality of grammars and their length lead to the minimum description length (MDL) principle by Rissanen [Rissanen, 1989]. This principle can be interpreted as follows: When comparing two different grammars, the bigger one might be able to interpret every output but it is not likely to generalize well. The best theory within the MDL principle is the simplest one which adequately describes the observed data. The *quality* of a grammar can be expressed in terms of length of the grammar itself and the given observation $\mathbf{O}$. This can be formalized by

$$G = \underset{G' \in \mathcal{G}}{\operatorname{argmin}} |G'| + |\mathbf{O}|_{G'}, \tag{1}$$

where $\mathcal{G}$ denotes the set of all possible grammars G which describe the observation data. $|G'|$ is the shortest encoding of the grammar G and $|\mathbf{O}|_{G'}$ is the shortest encoding of the observation $\mathbf{O}$ with given knowledge of grammar G'. With respect to Shannon's source coding theorem this can be rewritten as

$$G = \underset{G' \in \mathcal{G}}{\operatorname{argmin}} |G'| - \log P(\mathbf{O}|G'), \tag{2}$$

so every coding scheme for observations can be interpreted as a stochastic grammar and vice versa. In the *multigram* coding scheme the grammar consists of a lexicon. Every word inside the lexicon is associated with a probability that determines the relative frequency of that word. The MDL principle of equation (2) can be refined by

$$G = \underset{G' \in \mathcal{G}}{\operatorname{argmin}} \sum_{w \in G'} |w|_{G'} + \sum_{o \in O} |o|_{G'}, \tag{3}$$

where $|x|_{G'}$ is the description length of x using grammar G'.

Assuming that the codewords w are chosen to minimize the total description length, the codeword length $l(w)$ is related to the apriori probability of w by $l(w) = -\log P(w)$, so the coding system defines a stochastic language model. The probability of an observation sequence $\mathbf{O}$ under the grammar G is

$$\begin{aligned} P_G(\mathbf{O}) &= \sum_n P_G(n) \sum_{w_1 \ldots w_n = \mathbf{O}} P_G(w_1) \cdots P_G(w_n) \\ &\approx \sum_n \sum_{w_1 \ldots w_n = \mathbf{O}} P_G(w_1) \cdots P_G(w_n) \end{aligned} \tag{4}$$

Here the probability of $\mathbf{O}$ is given by summarization over the probabilities of all possible segmentations of $\mathbf{O}$ or in the context of codes over all possible representations of $\mathbf{O}$. The factor $P_G(n)$ describes the probability for a segmentation in n segments using this grammar and will be ignored during the rest of this paper. This kind of stochastic language model is called a *multigram model*. Multigrams reflect statistical dependencies within a sequence of letters by assigning a probability $P(w)$ to a variable length block w. When thinking in terms of observation of letters in an English text, the probability of $P(\texttt{the})$ should be larger than $P(\texttt{t}) \cdot P(\texttt{h}) \cdot P(\texttt{e})$. The modeling power of this multigrams can be greatly influenced by the maximum length of w. By increasing the length, the number of parameters increases exponentially, so there is a drawback between accuracy and the robustness in parameter estimation within this model.

The multigram is a finite-state model, it has only a finite memory of previous events which is restricted by the maximal length of codewords. Other well known finite-state models like n-gram models and HMMs are superior to the multigrams in the manner of modeling special sequences of observations, but multigrams tend to have a much smaller representation of the input data, which becomes obvious when comparing the number of parameters describing one word within the multigram framework and using n-grams.

As reflected above the maximization of equation (4) is equivalent of minimizing the description length of the underlying grammar. The maximization is done using a variant of the EM algorithm, which is equivalent to a Baum-Welch procedure. The expectation step consists of estimating the forward and backward variables

$$\alpha_i(\mathbf{O}) = \sum_{j=1}^{i-1} \alpha_j(\mathbf{O}) \sum_{w = o_{j+1} \ldots o_i \in G} P_G(w) \tag{5}$$

$$\beta_i(\mathbf{O}) = \sum_{j=i+1}^{l} \beta_j(\mathbf{O}) \sum_{w = o_{i+1} \ldots o_j \in G} P_G(w). \tag{6}$$

The probability of observing w spanning a region $o_a \ldots o_b$ is defined as

$$P_G(a \xrightarrow{w} b|\mathbf{O}) = \frac{\alpha_a(\mathbf{O}) P_G(w) \beta_b(\mathbf{O})}{P_G(\mathbf{O})} \tag{7}$$

The maximization step optimizes probabilities by normalizing the expected counts of parameters under the given lexicon G.

In [Deligne and Bimbot, 1995] the training process was started with all multigrams of a given maximum length which occur inside the training material. Due to the large number of parameters another method was proposed in [Marcken, 1996] which tries to reduce the number of parameters by starting with the simplest lexicon, where each multigram has a length of 1 and within each iteration altering the lexicon by adding new parameters and deleting obsolete ones, a method which will be used inside this paper.

3. BASE LINE SYSTEM

Our base line system for language identification consists of a two step process:

1. Extraction of language independent observation units which can be either codebook classes, phonemes or fenons.
2. Language dependent phonotactic modeling using n-gram models with $n = 1, 2, 3$ together with either discriminative [Ohler et al., 1999, Warnke et al., 1999] or usual interpolation schemes [Schukat-Talamazzini et al., 1997].

In the current system only phonotactic knowledge and no explicit knowledge on acoustic differences between languages is used. The stochastic framework is described as follows [see Harbeck et al., 1998]: The classification of an observation $\mathbf{X}$ is done selecting the language which yields the maximum a posteriori probability according to

$$\mathcal{LS}^* = \underset{\mathcal{LS}_j}{\operatorname{argmax}} P(\mathcal{LS}_j|\mathbf{X}) = \frac{P(\mathbf{X}|\mathcal{LS}_j)P(\mathcal{LS}_j)}{P(\mathbf{X})} \quad (8)$$

The idea is that speech is a sequence of unknown segments s_j like phonemes where every segments consists of a sequence of features vectors $\mathbf{x}_{s_j} \ldots \mathbf{x}_{s_{j+1}-1}$ which can be expressed as

$$\begin{aligned} P(\mathbf{X}|\mathcal{LS}_i) &= \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) P_{\mathcal{LS}_i}(\mathbf{X}|\mathbf{S}) \\ &= \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) \prod_{j=1}^{|\mathbf{S}|} P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}| \\ &\qquad \mathbf{x}_0, \ldots \mathbf{x}_{s_j-1}, s_j) \end{aligned} \quad (9)$$

$P_{\mathcal{LS}_i}(\mathbf{S})$ represents the phonotactic model, $P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|\mathbf{x}_0, \ldots \mathbf{x}_{s_j-1}, s_j)$ the probability for observing sequence $\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}$ within the segment s_j, which can approximated by

$$\begin{aligned} &P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|\mathbf{x}_0, \ldots \mathbf{x}_{s_j-1}, s_j) \\ &\quad \approx \quad P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|s_j). \end{aligned} \quad (10)$$

Corresponding to hidden Markov models we can formulate equation (9) as

$$\begin{aligned} P_{\mathcal{LS}_i}(\mathbf{X}) &= \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) \prod_{j=1}^{|\mathbf{S}|} P_{\mathcal{LS}_i}(\mathbf{x}_{s_j}..\mathbf{x}_{s_{j+1}-1}|s_j) \\ &= \sum_{s_1 \ldots s_{|\mathbf{S}|}} a_{s_1} \cdot a_{s_1,s_2} \cdot \ldots a_{s_1,s_2,\ldots,s_g} \cdot \\ &\qquad \prod_{l=g+1}^{|\mathbf{S}|} a_{s_{l-g}..s_l} \prod_{j=1}^{|\mathbf{S}|} P_{\mathcal{LS}_i}(\mathbf{x}_{s_j}..\mathbf{x}_{s_{j+1}-1}|s_j) \end{aligned} \quad (11)$$

where $a_{s_{l-g},\ldots,s_l}$ represents the conditional probability $P(s_l|s_{l-g}, \ldots, s_{l-1})$. The parameter g describes the order of statistical dependency, setting $g = 1$ will result in a normal HMM model. Every state within this model consumes not only one observation but a variable number of observations.
One of the big problems of this complex modeling structure is the initialization. To restrict the number of parameters inside the system we allow only segments with fixed length which we call the *codebook approach*. So equation (11) can be simplified to

$$P_{\mathcal{LS}_i}(\mathbf{X}) \approx \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) \prod_{j=1}^{n} P_{\mathcal{LS}_i}(\mathbf{x}_j|s_j). \quad (12)$$

When $P_{\mathcal{LS}_i}(\mathbf{S})$ is just a unigram model, this can be interpreted as a Gaussian mixture model. In our *codebook approach* the observation probability in equation (12) $P_{\mathcal{LS}_i}(\mathbf{x}_j|s_j)$ is approximated using a language independent observation probability function, which can be estimated by means of the standard LBG algorithm [Linde et al., 1980].

4. USING MULTIGRAMS FOR LANGUAGE IDENTIFICATION

In this section we describe two different kind of applications for multigrams inside our base line system.

4.1. Replacement for Language Models

The phonotactic model $P_{\mathcal{LS}_i}(\mathbf{S})$ is normally modeled by a stochastic n-gram language model and will be replaced by our multigram model with the codebook symbols as observations. Instead of calculating the probability of all possible segmentations as indicated in equation (4) only the probability of the best segmentation $s_1^* \ldots s_n^*$ is used

$$P_{\mathcal{LS}_i}(\mathbf{S}) = P_{\mathcal{LS}_i}(s_1^*) \cdots P_{\mathcal{LS}_i}(s_n^*) \quad (13)$$

4.2. Building a Fenon recognizer

In our opinion there are two major problems when using codebook classes for language identification:

- Codebook segments do not represent phonemes so phonotactic modeling based on codebook classes is not regular
- Codebook classes are very close inside the feature space so there is a tendency for substitution among them during recognition

It makes sense to search for more phoneme equivalent and more robust segments. One method to do this is to search for acoustic homogenous regions. But phonemes are not necessarily homogenous inside feature space and every phoneme shows a special movement or trajectory inside the feature space [Deng, 1993] which is indicated by different codebook classes. Typically the multigram approach is used in applications for unsupervised lexicon acquisition. The observation consists of letters where the word boundaries are not available, and the task is to find regular words inside the observation. Instead of letters we observe codebook classes, and instead of searching for words we are looking for sequences of codebook classes which are hopefully similar to phonemes.
The construction of the *fenon approach* is done with the following steps:

1. Train the codebook quantizer using LBG
2. Build the multigram language model using the quantized training material as observation
3. Estimate the most probable segmentation of the training material using the multigram model
4. Choose a subset of segments inside the best segmentation as fenons
5. Label the different fenons and use this as the new transcription
6. Train an HMM based recognizer on the new transcription
7. Use the fenon recognizer to extract the best fenons on the same training data, or if available on a disjunct training material
8. Train language specific phonotactic language models based on the output of the fenon recognizer

Like inside the *codebook approach* the acoustic frontend in this version is language independent and might be extended to language dependent models in the future. Only the phonotactic frontend represents language specific knowledge. The fenons do not have to represent only phonemes but are also able to represent common words like functional words which occur very often inside the training corpus.

5. EXPERIMENTS

In our experiments we used the languages German and English of the OGI corpus. As training set the training plus as validation annotated utterances are used (1 hour 20 minutes per language). As test either the test material annotated utterances (30 minutes per language) or the official NIST database was used (20 minutes per language). For comparison we evaluated in our first experiment the standard *codebook approach* and also used a supervised trained phoneme recognizer for language identification.

Method	*OGI test set*		*NIST test set*	
	10	30	10	30
Codebook approach	79	81	84	90
Phoneme approach	84	91	86	98
Multigrams	73	84	82	90
Fenons	76	87	87	98

Table 1. Recognition rates of language identification using different approaches for two languages on the OGI corpus evaluated on 10 and 30 seconds of speech.

As shown in table 1, the phoneme recognizer is the best on both sets when observing 30 second utterances. Comparing only the unsupervised trained approaches, the use of the fenon recognizer reduces the error rate of the *codebook approach* by 30 percent on the OGI test set and by 80 percent on the NIST test set which was even as good as using a supervised trained phoneme recognizer. When comparing the recognition rates on the 10 second utterances, the *codebook approach* is better than the fenon recognizer only on the OGI test set. So the use of fenons or phonemes seems to work especially on longer sentences. When the *multigram model* replaces the standard n-gram model the recognition rates drops down significantly on the 10 second sentences. On the 30 second sentences of the OGI test set the use of multigrams is better than using n-grams. One reason might be the artificial boundaries which are inserted into the observations when splitting the utterances into 10 second utterances. Also, there is no method to prevent over-adaptation to the training data as it is done inside the n-gram models.

6. CONCLUSION AND OUTLOOK

In this paper two new methods were proposed which are based on the information theoretic multigram models. These multigrams are developed to get a model for building a lexicon from scratch similiar to language acquisition. Using these models as a replacement for standard n-gram models does not improve the recognition. But it might be promising to combine both modeling schemes e.g. inside a neural network or train them using discriminative methods in the future.
Nevertheless, the use of multigrams for finding semiphonemes or fenons is quite promising as it increases recognition rate on the used test corpora, especially when observing long sentences and it is also as good as the supervised phoneme recognizer.

References

F. Bimbot and B.S. Atal. An Evaluation of Temporal Decomposition. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 1089–192, Genevo, Italien, September 1991.

N.A. Chomsky. *The Logical Structure of Linguistic Theory.* Plenum Press, New York, 1955.

S. Deligne and F. Bimbot. Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 169–172, Detroit, USA, 1995.

L. Deng. A stochastic model of speech incorporating hierarchical non-stionarity. *TIEEE*, 1(4):471–474, 1993.

S. Harbeck, E. Nöth, and H. Niemann. Multilingual Speech Recognition. In *Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pages 9–15, Pilsen, April 1997. University of West Bohemia.

S. Harbeck, E. Nöth, and H. Niemann. Multilingual Speech Recognition in the Context of Multilingual Information Retrieval Dialogues. In *Proc. of the Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, pages 375–380, Brno, September 1998. Masaryk University.

Y. Linde, A. Buzo, and R.M. Gray. An Algorithm for Vector Quantizer Design. *IEEE Trans. on Communications*, 28(1):84–95, 1980.

C.D. Marcken. *Unsupervised Language Acquisition.* PhD thesis, MIT, 1996.

U. Ohler, S. Harbeck, and H. Niemann. Discriminative training of language model classifiers. In *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, 1999.

J. Rissanen. *Stochastic complexity in statistical inquiry.* Singapure World Scientific, 1989.

E.G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. Interpolation of maximum likelihood predictors in stochastic language modeling. In *Proc. European Conf. on Speech Communication and Technology*, pages 2731–2734, Rhodos, Greece, 1997.

V. Warnke, S. Harbeck, E. Nöth, H. Niemann, and M. Levit. Discriminative estimation of interpolation parameters for language model classifiers. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 525–528, Phoenix, USA, 1999.

M.A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44, Januar 1996.