PHONETIC-CONTEXT MAPPING IN LANGUAGE IDENTIFICATION

Jiří Navrátil and Werner Zühlke

Department of Communication and Measurement
Technical University of Ilmenau, P.O.Box 100565, 98684 Ilmenau, Germany
e-mail: jiri.navratil@e-technik.tu-ilmenau.de

ABSTRACT

This paper deals with the problem of exploiting information from a wide phonetic context for the purpose of language identification. Two approaches to language modeling are presented here: 1) modified bigrams with a context-mapping matrix and 2) language models based on binary decision trees. Both models were incorporated in a phonotactic language identifier with a double-bigram decoding architecture and were shown to consistently improve the performance of standard bigrams. Measured on the NIST'95 evaluation set, the described system outperforms the state-of-the-art phonotactic components and is, at the same time, computationally less expensive.

1. INTRODUCTION

Automatic language identification (ALI) is a task of recognizing the language from a spoken test sentence. The ability of machines to distinguish between different languages becomes important with the trend in globalizing communication technology and providing wide multilingual services.

Besides other solutions for ALI based on prosody modeling, as well as phonetic acoustic features, there is an efficient way to describe a language in a discriminative way - by means of statistical modeling of phonetic chains (phonotactics). Several contributions were published dealing with the use of phone n-grams, particularly bigrams, which were shown to be suitable for identifying languages [2], [3], [4].

Although bigrams have proved to be efficient models, a wider phonetic context seems to be appropriate for acquiring language-relevant information. By introducing trigrams, i.e. second-order statistics, the performance of the phonotactic language models could be further improved. This, however, is faced with the general problem of lacking robustness due to sparse speech data. Moreover, as the probabilities are estimated from phonetic sequences decoded by a phone-recognizer that changes the original phonotactic properties of the language, a text-based estimation by means of pronounciation lexica and large text corpora is not feasible.

In this contribution two improved modeling methods will be described that allow acquiring information from a wider phonetic context than that of standard bigrams without increasing the estimation costs. In section 2. a simple algorithm for mapping the context of two preceding phones by means of a selection matrix is introduced. A more general approach to exploiting context information - by means of binary-tree language models - is presented in section 3. The subsequent sections detail the baseline ALI-system, the database and the phone-recognizer and give the experimental results obtained with both approaches.

2. APPROACH A: MAPPING WITH A SELECTION MATRIX

In general, the prior probability of a phone-sequence $\overline{a} = a_1, ..., a_T$ representing the spoken utterance, given a language model L_i, is calculated as

$$\Pr(\overline{a} \mid L_i) = \prod_{t=1}^{T} \Pr(a_t \mid a_{t-1}, ..., a_2, a_1, L_i).$$

The bigram is based on the approximation

$$\Pr(a_t \mid a_{t-1}, ..., a_2, a_1) \simeq \Pr(a_t \mid a_{t-1}),$$

i.e. all possible histories of the phone a_t are mapped into A equivalence classes (A being the size of the phone repertoire) each unifying those histories ending with the phone a_{t-1}. Although useful, the bigram approximation discards all statistical information of higher than the first order.

In order to exploit a wider context than that of standard bigrams while not increasing the estimation costs, a special selection function $S\{.\}$ is proposed that takes the history of two preceding phones into account and maps it into a manifold of equivalence classes of the size A. Thus the number of parameters to be estimated is reduced from A^3 (for trigrams) to A^2 (as for bigrams):

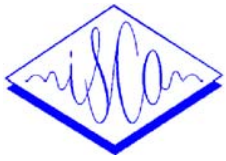
$$\Pr(a_t \mid x) = \Pr(a_t \mid S\{a_{t-1}, a_{t-2}\}).$$

Obviously, during this process a part of the 2nd-order statistical information gets lost. Determining S is the crucial problem in terms of minimizing the information loss. Therefore it is reasonable to take the phone-pair probabilities into account when designing the mapping rules S.

For the proposed algorithm two criteria were considered: 1) the resulting probability distribution $\Pr(c_k)$ of the new equivalence classes $c_1, ..., c_A$ should be nearly uniform and 2) the equivalence classes should have a comparable number of phone-pairs mapped in. The first criterion can be interpreted as maximizing the overall entropy of "utilization" of the equivalence classes, i.e. the average mapping rate should be equally distributed among the classes. As the phone-pair probabilities considered here represent the global occurences in all languages, the second requirement prevents one single language being preffered by the mapping function in cases when certain phone-pairs are very frequent uniquely in this language.

Naturally, the optimization task above may be solved in many different ways. The following algorithm represents a simple way to achieve this.

Fig. 1 illustrates the principle of the algorithm: a sequence of all A^2 possible phone-pairs h ordered according to their global occurence probability serves as the basis for deriving the equivalence classes. Typically, the sorted probabilities roughly fit an exponentially falling curve. In order to achieve the A^2-to-A reduction, the ordinate is folded up to the range $1...A$ in the manner depicted in the figure. It can be seen that the phone-pair probabilities in the columns $1...A$ sum up to an approximately uniform distribution thereby meeting the first criterion. Each of the columns $1...A$ represents a new equivalence class, and the phone-pairs assigned to it by folding are the class members. There are exactly A members assigned to each equivalence class, hence the second criterion mentioned above is satisfied as well.



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

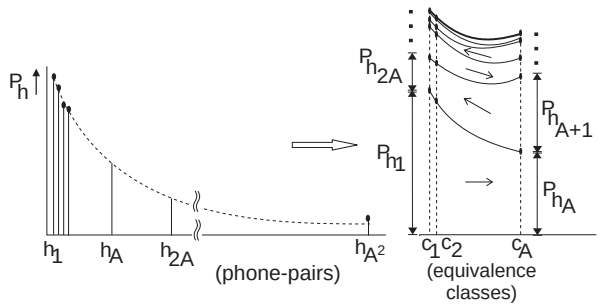


Figure 1. Ordered sequence of phone-pairs and the folding scheme

Based on this explanation the algorithm for generating the mapping rules can be formulated in four steps as follows:

1. Estimate the global phone-pair probabilities $\Pr(a_i, a_j)$ over all languages for $a_i, a_j \in \mathcal{A}$ ($\mathcal{A}$ being the phone repertoire of the size A).
2. Sort the probabilities in descending order. Let the elements of the ordered sequence be denoted $h_1, h_2, ..., h_{A^2}$.
3. Assign to the equivalence class c_i, $1 \leq i \leq A$ the following elements

$$c_i = \{h_i, h_{k \cdot A+i}, h_{k \cdot A-i+1}\}$$
$$\text{for } k = 2:2:A-1 \text{ and } A \text{ odd}$$
$$c_i = \{h_i, h_{k \cdot A+i}, h_{k \cdot A-i+1}, h_{A^2-i+1}\}$$
$$\text{for } k = 2:2:A-2 \text{ and } A \text{ even}$$

4. Fill the $A \times A$ selection matrix S so that each element S_{ij} contains the index of the equivalence class whose member the phone-pair $\{a_i, a_j\}$ is.

Although the sorted sequences seem to build nearly exponential descending curves, it is clear that the resulting shape of the class probability distribution depends on the actual language task, and its full uniformity cannot be guaranteed. However, as the experimental results will show, the described algorithm is an efficient - even though suboptimal - way to obtain a good mapping.

Once the selection matrix is generated, a set of modified bigram models can be estimated with their left context being transformed by means of S:

$$\Pr(a_t \mid S\{a_{t-1}, a_{t-2}\}, L_i) \approx \frac{N_{a_t, S\{a_{t-1}, a_{t-2}\}}}{N_{S\{a_{t-1}, a_{t-2}\}}}$$

(with N being the number of observations). Note that this new model set is assumed to be used as an addition to standard bigrams (see Section 4.).

3. APPROACH B: BINARY-TREE-BASED LANGUAGE MODELS

In [5] Bahl et al. introduced a special structure of language models based on binary decision trees for predicting words given a certain word history. Such models, when combined with word-n-grams, reduced the overall perplexity and proved to be feasible for natural language speech recognition.

Even though designed for large vocabulary speech recognition, the general structure of such models seems to be appropriate for modeling languages in terms of phonotactics as well. Obeying a minimum entropy rule and not being limited as to the length of history the tree-based models represent another promising way of phonetic-context acquisition.

A tree-model consists of nonterminal and terminal nodes (see Fig. 2). Each nonterminal node is connected with a

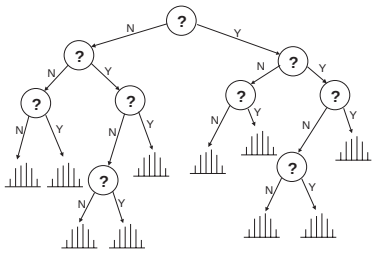


Figure 2. Example of a binary decision tree

binary question which leads to either one of two child-nodes. For answering the question a certain predictor (in this case a phone from the history) is compared with a node-dependent subset of phones. If the predictor belongs to the subset the result is positive, if not it is negative. When a terminal node (leaf) is reached, the probability of the phone a_t is obtained from the distribution. It is clear that different histories result in getting to different terminal nodes and different distributions. Thus, the context is exploited in a very flexible way.

In [5] a tree-growing algorithm was described that pursues minimizing the overall entropy of the phone distribution Y at each node. Regarding the expenditure, the advantage of this algorithm is the unsupervised construction of the tree on a training set without the need for linguistic experts. In the following, this algorithm is given (customized to phonotactics):

1. Let c be the current node of the tree. Initially c is the root.
2. For each predictor variable $X_i (i = 1, ..., m)$ find the subset $\mathbf{S}_i^c$ which minimizes the average conditional entropy at node c

$$\begin{aligned}\overline{H}_c(Y \mid \text{“}X_i \in \mathbf{S}_i^c?\text{”}) &= -\Pr(X_i \in \mathbf{S}_i^c \mid c) \sum_{j=1}^{A} \Pr(a_j \mid c, X_i \in \mathbf{S}_i^c) \\ &\cdot \log_2 \Pr(a_j \mid c, X_i \in \mathbf{S}_i^c) \\ &- \Pr(X_i \notin \mathbf{S}_i^c \mid c) \sum_{j=1}^{A} \Pr(a_j \mid c, X_i \notin \mathbf{S}_i^c) \\ &\cdot \log_2 \Pr(a_j \mid c, X_i \notin \mathbf{S}_i^c). \end{aligned} \quad (1)$$

3. Determine which of the m questions derived in Step 2 leads to the lowest entropy. Let this be question k, i.e.,

$$k = \arg\min_i \overline{H}_c(Y \mid \text{“}X_i \in \mathbf{S}_i^c?\text{”})$$

4. The reduction in entropy at node c due to question k is

$$R_c(k) = H_c(Y) - \overline{H}_c(Y \mid \text{“}X_k \in \mathbf{S}_k^c?\text{”}),$$

where

$$H_c(Y) = -\sum_{j=1}^{A} \Pr(a_j \mid c) \cdot \log_2 \Pr(a_j \mid c).$$

If this reduction is “significant,” store question k, create two descendant nodes, c_1 and c_2, pass the data corresponding to the conditions $X_k \in \mathbf{S}_k^c$ and $X_k \notin \mathbf{S}_k^c$, and repeat Steps 2-4 for each of the new nodes separately.

The principle of tree-growing is obvious: if the data at a node may be divided by a question in two sets having together a smaller entropy than the actual entropy of the undivided data, two new nodes are created. The entropy reduction is considered significant relative to some threshold.

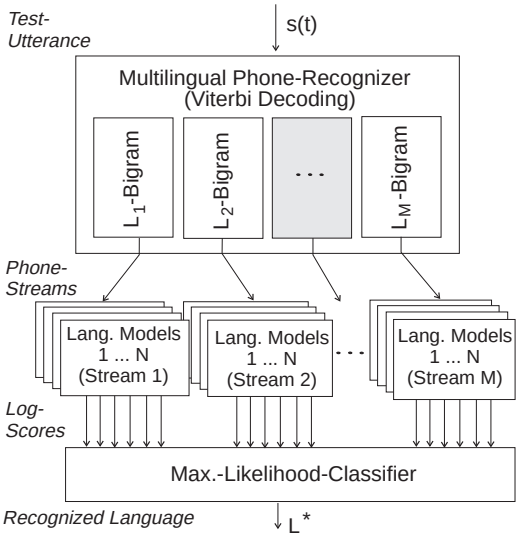


Figure 3. Baseline system overview

In order to determine the subset $\mathbf{S}_i^c$ in Step 2 a "greedy" algorithm was applied, as suggested in [5]. The search for $\mathbf{S}$ can be done through the following steps:

1) Let $\mathbf{S}$ be empty.

2) Insert into $\mathbf{S}$ the phone $a \in \mathcal{A}$ which leads to the greatest reduction in the average conditional entropy (1). If no $a \in \mathcal{A}$ leads to a reduction, make no insertion.

3) Delete from $\mathbf{S}$ any member a, if so doing leads to a reduction in the average conditional entropy.

4) If any insertions or deletions were made to $\mathbf{S}$, return to Step 2.

An example of a node-question could look like "$a_{t-1} \in \{/s/, /sh/, /f/\}$?", i.e. all phones that were preceded by the phones /s/,/sh/, or /f/ would be passed to the "yes"-node and vice versa. The optimal predictor in this case is a_{t-1}.

The essential parameter in the training algorithm is the significance threshold. Smaller thresholds will result in large trees with a great number of terminal nodes, whereas higher values will cause the tree to stop growing after few nodes. Further on, the number of predictors considered, i.e. the history length, is to be chosen. In our experiment both parameters were varied and their influence on the performance evaluated (see Section 5.2.).

4. BASELINE SYSTEM AND IDENTIFICATION

A phonotactic language identifier with a multilingual phone-recognizer and a double-bigram-decoding architecture [6] served as the baseline system for evaluating the new models (See Fig. 3). Here, during the Viterbi-decoding process, M (=6) language-dependent bigrams were used to weight the transitions between individual phones thus generating six phone-streams. With each stream an independent set of N language models is connected. Resulting scores are combined together and processed by a maximum-likelihood classifier. The bigrams used within the Viterbi-decoder were estimated on original transcriptions in six languages, whereas the language models were trained on the corresponding decoded phone-streams.

During the identification an incoming spoken utterance is "tokenized" into six phone-streams $\overline{a}^{(1)}, ..., \overline{a}^{(6)}$. Based on these, stream-dependent language bigram scores are calcualed for each language i and stream l:

$$S_{bi}(\overline{a}^{(l)} \mid L_i) = \frac{1}{T}\sum_{t=1}^{T} \log B(a_t^{(l)} \mid a_{t-1}^{(l)}, L_i)$$

where B denotes the interpolated bigram. Additionally, the scores for the selection-matrix and tree-based models are computed:

$$S_{sm}(\overline{a}^{(l)} \mid L_i) = \frac{1}{T}\sum_{t=1}^{T} \log B(a_t^{(l)} \mid S\{a_{t-1}^{(l)}, a_{t-2}^{(l)}\}, L_i),$$

$$S_{bt}(\overline{a}^{(l)} \mid L_i) = \frac{1}{T}\sum_{t=1}^{T} \log T(a_t^{(l)} \mid a_{t-1}^{(l)}, ..., a_{t-n}^{(l)}, L_i),$$

where T denotes the tree-model and n the number of predictors.

The stream-dependent score for a language L_i can be obtained by combining the individual model scores in an additive way:

$$S(\overline{a}^{(l)} \mid L_i) = S_{bi}(\overline{a}^{(l)} \mid L_i) + \alpha S_{sm}(\overline{a}^{(l)} \mid L_i) + \beta T(\overline{a}^{(l)} \mid L_i),$$

with α, β being empirical weights.

Finally, the classifier makes a maximum decision based on the total language scores as follows:

$$L^* = \arg \max_{1 \le i \le N} \sum_{l=1}^{M} S(\overline{a}^{(l)} \mid L_i)$$

5. IMPLEMENTATION

5.1. Database and the Phone Recognizer

Up to nine languages from the OGI Multi-Language Telephone Speech Corpus [7] were used for training and system development, and the NIST[1] test set from March '95 involving nine languages was taken for the system evaluations.

Twelve Mel-warped cepstral coefficients, energy as well as their first derivatives, were extracted from the signal waveforms, and the cepstral-mean substraction was carried out to suppress channel-dependent feature components.

For the phone-decoder an HMM-based phonetic recognizer was designed by means of the HTK software V2.0. The usual tri-state left-to-right model architecture for each individual HMM applied. 54 selected phonetic plus 7 non-speech HMM's (context-insensitive) were trained on speech signals in six languages, for which manually labelled and segmented transcriptions were available. A total number of 180 "stories-before-tone" (each 45 seconds long) served as the data to train the HMM parameters. The training conditions for the bigrams used within the Viterbi-decoder were described in [6].

Further on, 50+10 utterances (45s-stories) in each of the nine languages were decoded by the six-way phonetic recognizer and the resulting sequences served as data for training the language models as well as for tuning the system paramaters (α and β). These data did not overlap with the set used for phonetic training.

The NIST test set for each language consists of twenty 45-second phone calls (spontaneous monologues) and ca. eighty 10-second excerpts of them as specified in the NIST guidelines.

5.2. Training the tree-models

In order to investigate the influence of the tree-growing parameters - the significance threshold and the number of predictors - a variety of model configurations were evaluated. In addition to the significance value, another stopping criterion was implemented based on the distribution robustness. A node became terminal if the number of observations passed fell below a certain amount (≈ 900). This step proved to perform better than smoothing the distributions with the higher nodal distributions as proposed in [5].

[1] National Institute of Standards and Technology

Threshold	0.004	0.008	0.01	0.02	0.03	0.04
Avg.nodes	85	80	75	70	65	50

Table 1. Influence of the significance threshold on the tree size

Configuration	Error Rate	
	10s	45s
Baseline System	18.4%	5.0%
- with SM-Bigrams	14.7%	5.0%
- with Tree models	13.6%	2.5%
- with both	12.8%	3.3%

Table 2. Error rates on 10/45s utterances in the six-language-task (NIST'95)

Table 1 shows varying values of the entropy reduction threshold and the average number of resulting nodes. No significant variances were observed among the individual languages. The best performance was achieved with an average tree size of 75...80 nodes. Further on, the number of predictors was varied between 2, 3 and 4, i.e. three, four and five immediately neighboring phones in the sequence were modeled respectively. Here, the best results were measured when using three predictors. Typically, the gross structure of the tree was determined mainly by the predictor 1 (a_{t-1}), and predictors 2 and 3 seemed to be chosen more frequently in the lower nodes for rather detailed decisions.

6. EXPERIMENTS

Performance of the proposed system was tested using a closed set of six and nine languages. The efficiency of the selection matrix bigram (SM-model) and the binary tree-based model (BT-model) was examined by comparison to the baseline system.

6.1. Six-Language-Task

The following languages were taken for evaluation in the six-language-task: English, German, Hindi, Japanese, Mandarin and Spanish[2].

Table 2 shows the error rates for the baseline system and the resulting performance when either of the new models is added to the language models individually as well as when added in combination.

Both models achieved a consistent improvement whereby the BT-model seemed to work slightly better for both test lengths 10s and 45s - a reduction from 18.4% to 13.6% and 5.0% to 2.5% could be achieved for 10s- and 45s-utterances respectively. Adding both the BT-models and the SM-bigrams to the baseline system further reduced the error rate to 12.6% for the 10s-utterances, however, resulted in an increased error rate for longer utterances relative to the BT-model alone.

6.2. Nine-Language-Task

Comparable behavior of the new models can be seen in Table 3 for the nine-language-task in which the six languages listed above plus three other languages (French, Tamil and Vietnamese) were involved. In this case the combined models brought an overall improvement from 27.6% to 22.6% and 13.3% to 9.4% error rate. For longer test utterances the BT-model alone outperfomed the combination of both models, similar to the six-language-task.

7. DISCUSSION

Results obtained in the experiments clearly prove the efficiency of the described models. As expected, acquiring a wider phonetic horizon contributes to a better performance of standard bigrams in the phonotactic approach to ALI.

[2] Also chosen in the NIST evaluations in March '94

Configuration	Error Rate	
	10s	45s
Baseline System	27.6%	13.3%
- with SM-Bigrams	24.6%	12.2%
- with Tree models	24.0%	8.9%
- with both	22.6%	9.4%

Table 3. Error rates on 10/45s utterances in the nine-language-task (NIST'95)

Although the relatively simply-designed selection-matrix-based bigrams seem to be inferior to the more sophisticated tree-based models, their surprising performance in the 10s-tests uncovers the potential of the context information that remains unused by the standard bigrams.

The algorithm described in Section 2. designs a square-shaped selection matrix which automatically partitions the context into 54 equivalence classes. However, the optimal structure of the tree models resulting from the experiments makes obvious that more classes can be distinguished for a better performance. In this concern, the tree-based technique represents a powerful and flexible method to exploit the contextual information from the given training data.

While there was only one additional access operation per phone necessary to map the context for the SM-model, getting the phone probability of the tree model needed an average of five decisions per phone. With regard to the overall system, the additional models did not considerably increase the computational costs.

Measured on the NIST'95 data, the described system outperforms comparable state-of-the-art phonotactic systems [3][4]. Future research directions should address the question of how to combine phonotactic modeling with other approaches, e.g. prosody and acoustics, in order to obtain a more general system for robust ALI.

REFERENCES

[1] Y.K. Muthusamy, E. Barnard and R. A. Cole, "Automatic Language Identification: A Review/Tutorial," IEEE Signal Processing Magazine, October 1994.

[2] T.J. Hazen, V.W. Zue, "Recent improvements in an approach to segment-based automatic language identification," Proc. of the 1994 Int. Conf. on Spoken Language Processing, Yokohama, September, 1994, pp. 1883-1886.

[3] Y. Yan, E. Barnard, R. Cole, "Development of an approach to automatic language identification based on phone recognition," Computer Speech and Language, Vol. 10, No. 1, January 1996, pp. 37-54.

[4] M.A. Zissman, "Comparison of four approaches to automatic language identification," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, January 1996, pp. 31-44.

[5] L.R. Bahl, P.F. Brown, P.V. De Souza, R.L. Mercer, "A tree-based statistical language model for natural language speech recognition," IEEE Trans. on Acoustic, Speech, and Signal Processing, Vol. 37, No. 7, July 1989, pp. 1001-8.

[6] J. Navrátil, W. Zühlke, "Double-bigram decoding in phonotactic language identification," Proc. ICASSP-97, Munich, Germany, Vol. II, pp. 1115-8.

[7] Y.K. Muthusamy, R.A. Cole, B.T. Oshika, "The OGI multi-language telephone speech corpus," Proc. of the International Conference on Spoken Language Processing, Banff, Alberta, Oct. 12-16, 1992.