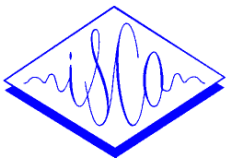

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

MULTI-LINGUAL TESTING OF A SELF-LEARNING APPROACH TO PHONEMIC TRANSCRIPTION OF ORTHOGRAPHY

Ove Andersen, Paul Dalsgaard
{oa,pd}@cpk.auc.dk
Center for PersonKommunikation, Aalborg University, Denmark

ABSTRACT

A Self-Learning system for Grapheme to Phoneme conversion is described and tested.

The system acquires the knowledge needed for grapheme-to-phoneme conversion from a training session in which a large number of pairs of grapheme strings and their corresponding (manually verified) phonemic transcription strings are presented to the system. The result from the training is a stochastic decision tree in which statistics - as given in the training material - about corresponding graphemes and phonemes are stored for later retrieval.

The system is tested on a number of European languages and results from three tests are reported.

In the first test, which concerns proper names, only the most probable phoneme candidate at each leaf of the tree is utilised. The second and the third test, both using a database of ordinary words, aims at analysing phoneme and word accuracies resulting from using N-Best phonemes at each leaf and from introducing phonotactic information, respectively.

Using N-Best candidates in combination with phonotactic information show a phoneme and word accuracy of up to 88.5% and 46.6%, respectively.

10.21437/Eurospeech.1995-280

1. INTRODUCTION

As the telecommunication sector in Europe is being liberalised these years the competition between the companies is increasing. One of the key factors for success is expected to be the services that each company can offer to their customers. Due to competition it is important to ensure cost-efficient services i.e. automatic services. Many of the potential automatic services will need access to knowledge about pronunciations of names. This include applications such as voice dialling, reverse directory enquiry, and credit card validation.

This paper describes a technique which has important implications in the context of phonemic transcription of orthography in general. The underlying idea is to establish a methodology capable of deriving automatically - from a quality controlled database - the 'rules' behind the process of converting graphemes into phonemes. The proposed technique may be utilised in several situations as e.g.

- for speeding up the establishing of pronunciation lexicons and
- as a fall-back utility in reading machines when a given word can not be found in the lexicon.

Results from a multi-lingual testing of a self-learning approach to grapheme-to-phoneme transcription are given.

Part of the work reported on in this paper was initially carried out within the LRE project ONOMASTICA. The objective of ONOMASTICA was to produce multi-language quality controlled pronunciation lexicons of proper names in a machine-readable form and to develop grapheme-to-phoneme rules tailored for proper names. Another important objective of the project was to develop self-learning systems, which deduce the information for grapheme-to-phoneme conversion from a training database. Such automatic acquisition of knowledge enables e.g. non-linguists to set up the conversion tool on any database with a minimum of work. The basic requirement is that a representative database of transcribed words is available.

2. PREVIOUS WORK

One of the earliest and most well-know examples of self-learning systems for grapheme-to-phoneme conversion is probably the work by Sejnowski from '87, see [1]. The work was based on the use of MLP neural networks and used for transcribing ordinary English words. The approach demonstrated an average phoneme transcription accuracy of approximately 95%. More recently ARPA in the USA has funded work on letter-to-sound/sound-to-letter generation which combines a rule-based formalism with data-driven techniques, see [2]. Here the authors reported on a phoneme transcription accuracy of 91.7% for letter-to-sound generation for English words.

The first results from work on the ONOMASTICA project on the establishment of a self-learning system (termed SELEGRAPH) for grapheme-to-phoneme conversion was presented in [3].

Here the task was the transcription of proper names and ordinary Danish words. The approach taken was based on the use of an iterative Viterbi alignment of graphemes and phonemes and a stochastic decision tree. The technique gave phoneme transcription accuracies for proper names and ordinary words of 92.0% and 94.9%, respectively.

3. ARCHITECTURE AND TRAINING

SELEGRAPH comprises three modules as illustrated in Figure 1, an alignment module, a module for context encoding and a module for storing (during training) and retrieving (during production) information about grapheme and phoneme correspondences.

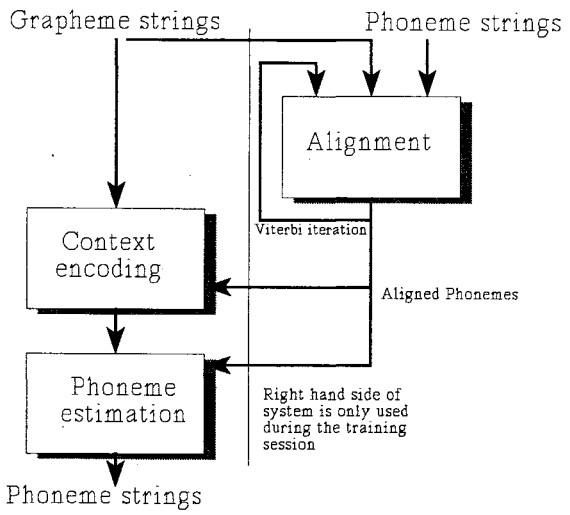

Figure 1. Architecture of SELEGRAPH

The training of the system addresses the following three problems:

- alignment of graphemes and phonemes
- calculation of contextual graphemic information
- storing of the stochastic information

which are handled in a three-step procedure.

3.1 Step 1 - Alignment

It is a basic requirement of the system that grapheme strings and their corresponding phoneme strings are of equal length in order to acquire the knowledge needed for grapheme-to-phoneme conversion.

The alignment of two strings are achieved by automatically inserting graphemic and phonemic nulls in the strings by means of an iterative Viterbi alignment algorithm. Details are given in [4].

3.2 Step 2 - Context Encoding

Aiming at high-accuracy, automatic learning of grapheme-to-phoneme conversion it is essential to take the context into consideration, and to decide how many characters to include from the right and from the left context of the current grapheme. The decision on the size of the context is guided by computing the mutual information of the graphemes in the context as proposed in [5]. Another important output of the computation of mutual information is the ordering in which the context is to be considered.

3.3 Step 3 - Phoneme Estimation

The SELEGRAPH system acquires the knowledge on grapheme-to-phoneme conversion by being exposed to a database of representative pairs of grapheme and phoneme strings. The information is stored in a tree structure which enables fast retrieval of data, easy handling of unseen grapheme sequences and minimum memory requirements.

The tree consists of leaves and branches. Each leaf records statistics about one grapheme G_j in a well defined context. The statistics is simply a list which, for each grapheme G_j, contains the number of occurrences n_i of each possible phoneme Φ_i as found in the specific context in the entire training database. Furthermore, each leaf has a pointer to the following leaf.

The path into the levels of the tree structure is defined by the ordering as found during step 2 of the training procedure i.e. graphemes with highest mutual information are considered first. Each level contains a number of leaves each of which corresponds to a grapheme as represented in the training database. The connection between a leaf at a certain level and the leaves at the preceding and following levels gives the graphemic context.

An example of part of a tree structure and the information stored at each leaf on the basis of the training session is shown in Figure 2.

It is noticed that the statistics for several phonemes may be stored at each leaf. This is further commented in section 5.1 below.

During testing the tree is searched until there is no further match.

4. MULTI-LINGUAL TESTING OF THE BASELINE SYSTEM

In the baseline system only the statistics of the most probable phoneme is stored at each leaf.

The work presented in [3] mainly focused on Danish proper names. This paper demonstrates that the technique is immediately applicable to proper names of other languages as well. The application of SELEGRAPH on the following languages - Danish, British-English, Norwegian, Italian and Spanish - gives a quantitative comparison of the complexity of transcribing proper names for these languages.

The results are given in Table I together with information on the material available for training and testing.

The resulting evaluation of the automatically derived transcriptions is carried out in a two step procedure.

First, the manually and the automatically given strings are aligned and secondly the number of agreements are registered. Correctly transcribed words are transcriptions without errors at the

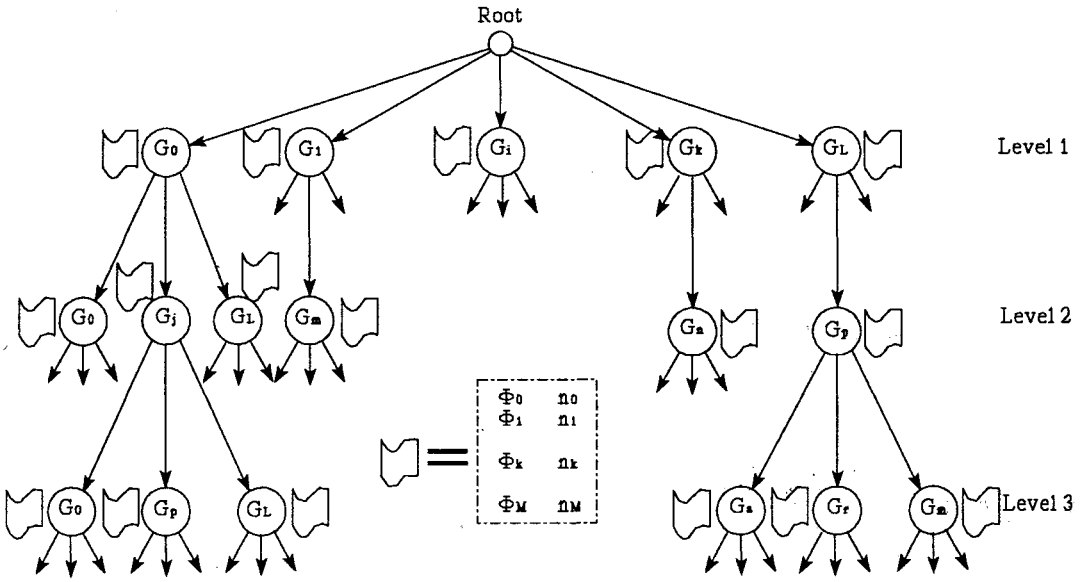

Figure 2. Structure of the stochastic decision tree

Table I. Number of manually transcribed proper names and phoneme transcription accuracies estimated on the basis of subsets of the total databases.

Lang.	Total # of words	# words training	# words testing	Phoneme accuracy
Danish	135.503	8.924	8.926	94.4%
English	137.721	15.000	10.000	93.9%
Norweg	73.329	10.000	10.000	92.5%
Italian	84.847	18.999	18.876	99.4%
Spanish	188.760	4.999	4.960	98.7%

phoneme level.

The results show a separation of the values of phoneme transcription accuracies into two groups. The Romanic and the Germanic language groups have average phoneme transcription accuracies of 99.1% and 93.6%, respectively.

5. EXTENDED SYSTEM TESTING

The following tests both use the publicly available American-English 'NETtalk' database [1], which contains approx. 20.000 ordinary words, half of which are used for training and the remaining for testing.

The tests are based on two extensions introduced into the baseline system and its postprocessing. Firstly, each leaf is allowed to store statistics for a number of N-Best phonemes (see section 5.1) and secondly, phonotactic information is used in the postprocessing giving the resulting estimated transcription (see section 5.2). The phonotactic information is given from the training database.

5.1 N-Best phonemes at each leaf

The aim of this test is to check the possible drop in transcription error rates which may be obtained by using N-Best phonemes instead of only one at each leaf. Observe, the number of alternative phonemes is likely to be smaller the deeper the level within the tree.

Table II gives phoneme and word error rates as a function of the number of N-Best phonemes retained per leaf.

It is important to notice that Table II gives the minimum error rates that can be obtained, as the results are only a registration of how often the correct phoneme is among the N-best. An automatic method for selecting a single estimate from the N-best is tested in the next section.

Table II. Phoneme and word error rates as a function of the number of phonemes represented at each leaf.

N Best	Phoneme	Word
1	11,7%	53,8%
2	7,6%	39,2%
3	6,8%	35,6%
4	6,6%	34,7%
5	6,6%	34,5%
All	6,5%	34,4%

It is observed that a reduction of approximately 42% and 35% in phoneme and word error rates, respectively, may be achieved by applying the three most likely phonemes per leaf rather than just the single most probable phoneme.

5.2 Utilising Phonotactic Information

In this test all phonemes are retained at each leaf in the decision tree, and the aim is to justify a possible increase in the transcription accuracies by applying phonotactic information in the postprocessing. To enable this, the training database is searched for occurrences of all possible phoneme pairs, i.e. 'bigram phonotactics'.

As the 'NETtalk' database contains 51 different phonemes there exists theoretically 2601 possible phoneme pair combinations. It is found that the (limited size) training material - here consisting of approximately 10.000 words - contains only 1047 different phoneme pairs.

Each possible combination may be represented by two numbers. The first representing the 'bigram probability' - the relative frequency of occurrence - which can be calculated directly from the training material. The second representing a 'pseudo probability' is a registration of whether a phoneme-pair is present - then assigned the probability '1' - or absent - then given the value '0' - in the training material.

These 'probabilities' are used during the Viterbi-based postprocessing in which the final phoneme transcription string for a given word is estimated.

Table III. Percentage phoneme and word error rates as a function of the window size and by using three methods for phonotactic parsing.

	Window 1-1-1			Window 3-1-3		
	A	B	C	A	B	C
Phoneme	17,7	17,5	17,4	11,7	11,6	11,5
Word	72,4	72,2	71,4	53,8	53,6	53,4

A: No phonotactic information used (corresponds to 1-Best); B: 'Bigram' phonotactics; C: 'Pseudo' phonotactics

The probability for a phonemic transcription Φ given the orthography G can be estimated as:

$$p(\Phi \mid G) = p(\Phi_{i,1} \mid g_{c,1})\ p(\Phi_{k,2} \mid \Phi_{i,1}) \quad \cdots \\ p(\Phi_{i,j} \mid g_{c,j})\ p(\Phi_{k,j+1} \mid \Phi_{i,j}) \cdots \\ p(\Phi_{i,L-1} \mid g_{c,L-1})\ p(\Phi_{k,L} \mid \Phi_{i,L-1})$$

where

- $p(\Phi_{i,j} \mid g_{c,j})$ is the probability - given the centre grapheme G_j of the j'th context $g_{c,j}$ - of transcribing G_j into phoneme $\Phi_{i,j}$; *i* indexing the phoneme $\Phi_{i,j}$ of the set $\Phi_{i,j} \in \{\Phi_1 \bullet \bullet \Phi_M\}$ and M being the number of phonemes of the language in question
- $p(\Phi_{k,j+1} \mid \Phi_{i,j})$ is the 'bigram or pseudo probability' of the transition of phoneme $\Phi_{i,j}$ into phoneme $\Phi_{k,j+1}$ and
- L is the number of graphemes in the word being transcribed.

The Viterbi search may be performed on the based of either 'bigram probabilities' or 'pseudo probabilities'.

By maximising the above expression, a final estimate of the pronunciation is obtained. The results from these tests are shown in Table III, which furthermore shows the importance of selecting a proper contextual window during transcription. It is e.g. seen that a 3-1-3 window - a context window of length seven graphemes - shows a substantially lower error rate than for a 1-1-1 window.

6. CONCLUSIONS

The paper has demonstrated that the established self-learning approach can be used without modifications of the original implementation on several European languages. Furthermore, the tests have shown that it is possible to increase the phoneme and word transcription accuracies by utilising an N-Best number of phonemes within the decision tree, and in applying phonotactic information.

However, even lower error rate at the word level can possibly be obtained by utilising phonotactic information which extends to more than just bigrams. This hypothesis has to be tested on training material which is larger than the NETtalk database. Data from the ONOMASTICA project - now available on CD-ROM - are an excellent basis for such further testing.

7. ACKNOWLEDGEMENTS

The work presented in this paper has partly been funded by the ONOMASTICA project and partly by CPK via funding from the Danish Technical Research Council.

8. REFERENCES

[1] T.J. Sejnowski and C.R. Rosenberg. "Parallel networks that learn to pronounce English text." Complex Systems, pp. 145 - 168, 1987.

[2] H.M. Meng, S. Seneff, and V.W. Zue. "Phonological parsing for reversible letter-to-sound /, sound-to-letter generation." In Proceedings of International Conference on Acoustics, Speech and Signal Processing, Adelaide, Australia, volume, pp. 1 - 4, September 1994.

[3] O. Andersen and P. Dalsgaard. "A self-learning approach to transcription of Danish proper names." In International Conference on Spoken Language Processing, Yokohama, Japan, pp. 1627 - 1630, 1994.

[4] P. Dalsgaard, O. Andersen and A.V. Hansen. "Theory and Application of two Approaches to Grapheme-to-Phoneme Conversion", Deliverable 4.7 from the ONOMASTICA project, CPK, May 1995.

[5] J.M. Lucassen and R.L. Mercer. "An information theoretic approach to the automatic determination of phonemic baseforms." In Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 42.5.1 - 42.5.4, 1984.