

SPE based selection of context dependent units for speech recognition

Matjaž Rodman[♦], Bojan Petek and Tom Brøndsted*

Interactive System Laboratory
Faculty of Natural Sciences and Engineering
University of Ljubljana
Snežniška 5, 1000 Ljubljana, Slovenia
matjaz.rodman@ntftex.uni-lj.si, bojan.petek@uni-lj.si

*Center for PersonKommunikation (CPK)
Institute of Electronic Systems
Aalborg University
Niels Jernes Vej 12, 9220 Aalborg, Denmark
tb@cpk.auc.dk

Abstract

Decision tree-based approach is a well known and frequently used method for tying states of the context dependent phone models since it is able to provide good models for contexts not encountered in the training data. In contrast to the other approaches, this method allows us to include expert linguistic knowledge into the system. Our research focused on the inclusion of standard generative theory by Chomsky & Halle (1968), called the SPE theory (the Sound Pattern of English), into the decision tree building process as expert linguistic knowledge. Our attempt was to "merge" the SpeechDat2 SAMPA label set, used for English and Slovenian languages, with the SPE. We created all possible natural groups of phones (SAMPA segments defined by a set of binary phonological features) for both languages and included them into a set of questions used in the process of creating the decision trees. Based on the decision tree constructed this way, we created an English and Slovenian speech recognition systems and tested both of them. Compared with the reference speech recognition system (Lindberg et al., 2000; Johansen et al., 2000) we got some promising results that encouraged us to continue this work and to perform further testing.

1. Introduction

Much of the phonetic variation in natural speech is due to contextual effects. In order to be able to accurately model variations in natural speech a careful choice of the units represented by each model is required. In large-vocabulary speech recognition systems, modelling of vocabulary words by subword units (phonemes or units derived from phonemes) is mandatory. For example, triphone models have been one of the most successful context dependent units because of their ability to model well the co-articulation effect. Yet if we create distinct models for all possible contexts, the number of models becomes very high. In practical applications of building speech recognition systems, there is often a conflicting desire to have a large number of models and model parameters in order to achieve high accuracy, whilst at the same time having limited and uneven training data in form of labelled utterances of a particular language (Young et al., 2000). In the case of triphone context dependent models, tying of HMM states gives us a possible solution of how to overcome this problem.

In our work we analysed the influence of the decision tree method on the acoustic modelling. We also analysed parameters that influence the decision tree building process and tested the proposed method based on the theory of naturalness (the theory that phonological segments cluster into "natural groups" defined by universal features), (Chomsky et al., 1968). We first examined this issue within the Slovenian language and then also addressed its portability to other languages.

2. Decision tree

When building large vocabulary cross-word triphone systems, unseen triphones are unavoidable. A limitation of the data-driven clustering procedure is that it does not deal

with triphones for which there are no examples in the training data. Decision tree based approach gives us a possibility to include expert linguistic knowledge into a procedure of creating acoustic models. This methodology provides appropriate models also for contexts that are not seen in the training data. Therefore, decision trees are used in speech recognition with large numbers of context dependent HMMs, to provide models for contexts not seen in the training data. Sharing data at the model level may not be the most appropriate method for models composed of distinct states (Odell, 1995). Sharing distributions at the state level allows for finer distinctions to be made between the models by allowing left and right contexts to be modelled separately.

2.1. Decision tree building process

A phonetic decision tree is a binary tree in which a yes/no phonetic question is attached to each node (Young et al., 2000). Initially, all states in a given item list (typically a specific phone state position) are placed at the root node of a tree. Depending on each answer, every node is successively split and this continues until the states have trickled down to leaf-nodes. All states in the same leaf node are then tied and trained from the same data.

The question at each node is chosen to (locally) maximise the likelihood of the training data (using a log likelihood criterion) and gives the best split of the node. This process is repeated until the increase in log likelihood falls below the specified threshold. As a final stage, the decrease in log likelihood is calculated for merging terminal nodes, which belong to different parents. Any pair of nodes for which this decrease is less than the threshold used to stop splitting is then merged (Young et al., 2000). The algorithm for building a decision tree is summarised in figure 1.

[♦] Socrates/Erasmus exchange student under the multilateral agreement UL D-IV-1/99-JM/Kc.

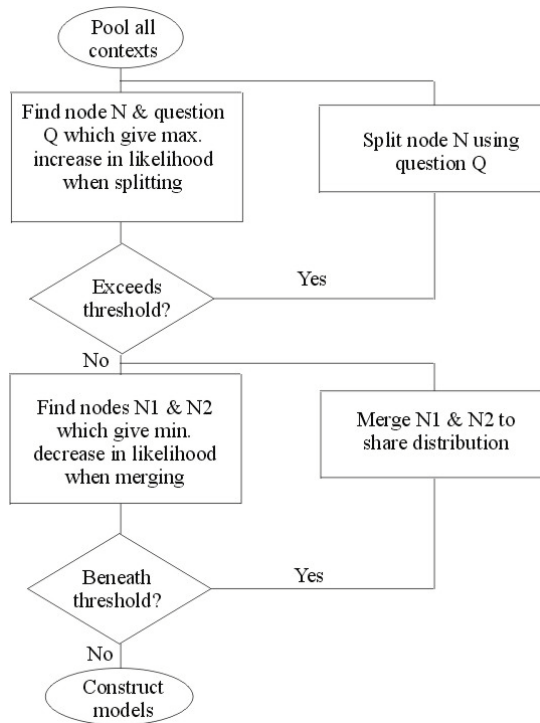


Figure 1. Algorithm for constructing decision tree (Odell, 1995)

Questions asked in the decision tree have a form:

QS "L_SL_Nasal" { m-,n-*,N-* }*

As an example, the command above defines the question “Is the left context a nasal?” where the group of nasals is represented by $\{m-*, n-*,N-*\}$. Only a finite set of questions can be used to divide each node. So questions have to be defined in a way that all possible natural groups of phonological segments are stated. That allows the incorporation of expert linguistic knowledge needed to predict contextual similarity when little or no data is available in order to determine which contexts are acoustically similar.

Decision tree building process has two stop criteria that determine how deep the tree will be. The first one is increase in the log likelihood that has to be achieved if node was split. In HTK (Young et al., 2000) it is defined with the command TB. The second one is the minimal occupation count that determines how many training data each node has to have. In HTK it is defined with the command RO.

3. SPE theory

Distinctive feature theory was introduced first by R. Jakobson. He set up twelve universal inherent feature classes. Chomsky and Halle took over Jakobson's idea and defined 22 universal feature classes, which according to the standard SPE theory are sufficient for analysing expression segments of any language into distinctive oppositions.

The idea of natural phonetic groups is based on the so-called Sound pattern of English theory, “SPE”, of

Chomsky & Halle (1968). By this theory an inventory of expression segments can be described in terms of a hierarchical tree structure where upper nodes represent major class features (like +/- vocalic, +/- consonantal) and lower nodes cavity features, manner of articulation etc., and terminal nodes represent phones. A phonetic representation of an utterance in a given language has by this theory the form of a two-dimensional matrix in which the rows are labelled by features of universal phonetics; the columns stand for the consecutive segments of the utterance generated; and the entries in the matrix determine the binary value (+/-) of each segment with respect to the universal features (Chomsky et al., 1968). A set of phonological segments (“phonemes”) sharing the same feature matrix and unequivocally defined by this matrix form a natural group. There are more degrees of naturalness. The SPE theory claims that one group is more natural than the other if the number of features defining it is smaller. The main natural groups (vowels, consonants, semi-vowels) are separated just by different values in major class features. Specific groups (e.g. back-vowels, plosives, nasals, labials) are defined by further features in the matrix and are consequently “less natural”. Groups of segments that cannot be defined by a feature matrix are not natural (e.g., the pseudo group: k, a, m, h).

3.1. The use of SPE on SpeechDat2 databases

The starting point of our distinctive features composition can be described as follows:

- We intended to use the SPE as a generally accepted standard theory of phonology and with as few modifications as possible.
- Most notably, we have tried to utilise the Chomsky & Halle decomposition of English segments (1968) as directly as possible.
- Finally, we have attempted to make as few changes to the SpeechDat2 label set as possible.

Hence, our starting point can be paraphrased as attempt to “merge” the SAMPA label set used in SpeechDat2 database with the SPE.

The SPE sets up a total number of twenty-two feature classes, which according to the standard theory are sufficient for analysing expression segments (phonemes) of any language into distinctive oppositions. For a distinctive feature composition of the segments of a specific language, not all 22 feature classes are utilised. For instance, the SPE-description of English segments (Chomsky et al., 1968) makes references only to 13 feature classes. The remaining 9 classes may be regarded as redundant or “irrelevant” to English.

The set of 15 features was sufficient to represent the set of Slovenian and English SAMPA symbols used in the SpeechDat2 database by the standard SPE theory. In general, we tried to preserve the original distinctive features used in the SPE. We had to, however, make some changes. In short, we replaced the feature vocalic with sonorant and syllabic, and added a feature front (Brøndsted, 1998). The feature +/- front is not within the set of 22 universal binary features defined in the SPE. However, the feature is needed additionally to +/-back because the SAMPA symbols include segments of a dubious phonological state, only specifiable with reference to three places of articulation: [-back, +front], [-back, -front], and [+back, -front].

3.2. Major Class Features

In standard generative phonology, the major class features sonorant, syllabic and consonantal are used to classify phonological segments into five major groups: vowels, non-syllabic liquids/nasals, syllabic liquids/nasals, glides, and obstruents. However, as the SAMPA segments defined for English and Slovenian do not include syllabic liquids/nasals, this in our case resulted in only four major groups (cf. table 1).

	Sonorant	Syllabic	Consonantal
Vowels	+	+	-
Glides	+	-	-
Syllabic Liquids and Nasals	+	+	+
Non-Syllabic Liquids and Nasals	+	-	+
Obstruents	-	-	+

Table 1: The main natural groups represented by major class features

3.3. The use of SPE on the Slovenian SpeechDat2 database

To create a distinctive feature composition table of the Slovenian SAMPA symbols used in SpeechDat2 we had to modify the phonetic transcriptions. In total, SpeechDat2 uses 46 SAMPA symbols in the Slovenian transcriptions. However, according to (Šuštaršič, 1999; Toporišič 2000) Slovenian only has 29 phonemes. Thus, 17 symbols must be considered allophonic variants. These allophones include certain composite pseudo segments (t_n, d_n, p_n, b_n, t_l, d_l) used along with the normal polyphonematic transcriptions (t n, d n ... etc.) in a way that appeared non-systematic to us. Consequently, we decided to change phonetic transcriptions in the database according to the following seven rules:

- Change string “t_n n” with two symbols “t n”
- Change string “d_n n” with two symbols “d n”
- Change string “p_n n” with two symbols “p n”
- Change string “b_n n” with two symbols “b n”
- Change string “t_l l” with two symbols “t l”
- Change string “d_l l” with two symbols “d l”
- Change symbol “W” with symbol “w”

This reduced the set of segments from 46 to 39. The resulting distinctive feature composition table of the Slovenian vowel and consonantal segments is shown in tables 2 and 3.

3.4. The use of SPE on the English SpeechDat2

Similarly we had to modify the transcriptions of the English SpeechDat2 database. The major problem was the monophonematic representation of diphthongs (as single phones). In SPE theory there are no phonological features differentiating diphthongs from monophthongs. This theory handles diphthongs with certain appropriate diphthongisation rules applied to the underlying representations (Chomsky et al., 1968). In order to provide a level of description conforming to the underlying

	i	i:	e	e:	ɛ	ɛ:	a	a:	u	u:	o	o:	@	@r
Sonor.	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Syllabic	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Conson.	-	-	-	-	-	-	-	-	-	-	-	-	-	-
High	+	+	-	-	-	-	-	+	+	-	-	-	-	-
Back	-	-	-	-	-	-	+	+	+	+	+	+	-	-
Front	+	+	+	+	+	-	-	-	-	-	-	-	-	-
Low	-	-	-	-	+	+	+	+	-	-	-	+	+	-
Round	-	-	-	-	-	-	-	+	+	+	+	+	-	-
Tense	-	+	-	+	-	+	-	+	-	+	-	+	-	-
Anterior														
Coronal														
Voice														
Cont.														
Nasal														
Strident														

Table 2: Distinctive feature composition of Slovenian vowel segments

	b	d	g	p	t	k	dZ	ts	tS	s	S	Z	Z	f	v
Sonor.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Conson	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
High	-	-	+	-	-	+	+	-	+	-	+	-	+	-	-
Back															
Front															
Low															
Round															
Tense															
Anterior	+	+	-	+	+	-	-	+	-	+	-	+	-	+	+
Coronal	-	+	-	-	+	-	+	+	+	+	+	+	+	+	-
Voice	+	+	+	-	-	-	+	-	-	-	-	+	+	-	+
Cont	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Strident	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+

Table 3: Distinctive feature composition of Slovenian consonantal segments representation presupposed by the SPE, the diphthongs were re-written according to the 8 rules:

	w	j	x	r	l	m	n	N
Sonor	+	+	+	+	+	+	+	+
Syllabic	-	-	-	-	-	-	-	-
Conson.	-	-	-	+	+	+	+	+
High	+	+	-	-	-	-	-	+
Back	+	-	-	-	-	-	-	-
Front	-	+	-	-	-	-	-	-
Low	-	-	+	-	-	-	-	-
Round	+	-	-	-	-	-	-	-
Tense	-	-	-	-	-	-	-	-
Anterior	-	-	-	-	+	+	+	-
Coronal	-	-	-	-	+	+	+	-
Voice				-	+	+	+	+
Cont.				+	+	+	-	-
Nasal				-	-	-	+	+
Strident				-	-	-	-	-

- Change symbol “eI” with phones “e” and “j”
- Change symbol “aI” with phones “{” and “j”
- Change symbol “OI” with phones “Q” and “j”
- Change symbol “@U” with phones “@” and “w”
- Change symbol “aU” with phones “{” and “w”
- Change symbol “I@” with phones “I” and “@”
- Change symbol “e@” with phones “e” and “@”
- Change symbol “U@” with phones “U” and “@”

The resulting distinctive feature composition of the English vowels and consonants are presented in tables 4 and 5.

	i	u	ɜ	O	A	I	U	e	{	Q	V	@
Sonor.	+	+	+	+	+	+	+	+	+	+	+	+
Syllabic	+	+	+	+	+	+	+	+	+	+	+	+
Conson.	-	-	-	-	-	-	-	-	-	-	-	-
High	+	+	-	-	-	+	+	-	-	-	-	-
Back	-	+	-	+	+	-	+	-	-	+	+	-
Front	+	-	-	-	-	+	-	+	+	-	-	-
Low	-	-	-	+	+	-	-	-	+	-	-	-
Round	-	+	+	+	-	-	+	-	-	+	-	-
Tense	+	+	+	+	-	-	-	-	-	-	-	-
Anterior												
Coronal												
Voice												
Cont.												
Nasal												
Strident												

Table 4: Distinctive feature composition of English vowel segments

3.5. Definition of natural groups

During the process of creating the decision tree, groups of phones are used to define questions that may be used in each node of the decision tree. This is the most important stage in the entire model-building procedure where expert phonological knowledge can be included (another one is the prior stage, where the actual set of phones to be used for segmentation and classification of the acoustic signal is established). For that reason, groups of phones for five languages - among these both Slovenian and English - were defined as a part of the COST 249 project. As the languages partly use the same phonemic label set (SAMPA), the groups are reusable across languages. Slovenian contributes with 45 groups and English with 17 groups. During the process of creating the decision tree, two questions are created from every group defined. One is about the left context and the other about the right one. On the basis of these definitions we created English and Slovenian reference recognition systems.

Our main goal was to create another two systems for both languages that would have phone groups defined on the basis of the SPE theory. Therefore we automatically generated all natural groups of phones from the distinctive feature compositions table set up for the two languages. This resulted in 174 natural groups for Slovenian and 171 for English. The groups were used to create the set of all possible questions to be included in the process of building the experimental SPE-based speech recognition systems.

	b	d	g	p	t	k	dZ	tS	s	S	z	Z	f	T	v	D
Sonor.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Conson	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
High	-	-	+	-	-	+	+	-	+	-	+	-	+	-	-	-
Back																
Front																
Low																
Round																
Tense																
Anterior	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+
Coronal	-	+	-	-	+	-	+	+	+	+	+	+	-	+	-	+
Voice	+	+	+	-	-	-	+	-	-	-	+	+	-	-	+	+
Cont	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Strident	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	+

	w	j	h	r	l	m	n	N
Sonor	+	+	+	+	+	+	+	+
Syllabic	-	-	-	-	-	-	-	-
Conson.	-	-	-	+	+	+	+	+
High	+	+	-	-	-	-	-	+
Back	+	-	-	-	-	-	-	-
Front	-	+	-	-	-	-	-	-
Low	-	-	+	-	-	-	-	-
Round	+	-	-	-	-	-	-	-
Tense								
Anterior	-	-	-	-	+	+	+	-
Coronal	-	-	-	-	+	+	-	-
Voice			-	+	+	+	+	+
Cont.			+	+	+	-	-	-
Nasal			-	-	-	+	+	+
Strident			-	-	-	-	-	-

Table 5: Distinctive feature composition of English consonantal segments

4. Importance of the order of questions for “unseen” contexts

We hypothesised a case of why it would be not advisable to create questions that would include all possible combinations of phonemes (including "unnatural" groups) and leave to the decision tree building process to chose the best ones by it's own criteria. This way the decision tree building process would pick up only the important questions (likely involving only "natural" groups) and leave out the irrelevant ones. The idea emerged because of the explanation in the HTK documentation considering the problem of how to build questions for a decision tree: “There is no harm in creating extra unnecessary questions, because those which are determined to be irrelevant to the data will be ignored” (Young et al., 2000). That would yield us the optimal decision tree for this particular system without including any linguistic knowledge. By this definition also the order of the questions in the file that HTK uses for creating a decision tree should have no effect on the structure of the

decision tree. But already the first experiment showed us that the order of questions in this file *does* matter.

When we changed the order of questions in the file also the structure of decision tree has changed. Considering how questions are chosen in the process of building decision tree, we got a possible explanation for this change. For example let's suppose that we are in the process of deciding how to cluster the centre state of the phone /m/. Let's assume that we have data only for the triphones $a-m+*$, $b-m+*$, $c-m+*$ and $d-m+*$ where * means any context. Suppose further that we have defined the questions $QS "L_context1" \{a-*, b-*, x-*$ and $QS "L_context2" \{a-*, b-*$ where the first one is a superset of the second one (including also the left context 'x'). The log-likelihood can only be calculated for data that is available for training. Therefore these two questions would cause the same increase in log likelihood if they were used for splitting the node because the left context $x-*$ does not appear in the training data. So if $L_context1$ was used, the middle state of the model with the left context x would be trained from the same data as middle states of the models with left contexts a and b ! Likewise, if $L_context2$ was used, the middle state of the model with the left context x would be trained from the same data as the middle states of the models with left contexts c and d so from different data as in the first case. Both situations are presented in figure 2. Increase in log-likelihood would be the same in both cases. Therefore, only the order of questions in the file where questions are defined or the procedure that defines which question to use, if more questions give the same increase in log likelihood, would decide from which data model with left context x was trained. This means that for the models with contexts not seen in the training data (like x here) the decision from which data they'll be trained would depend on the order of questions.

From this we concluded that the phone groups that are later transformed into questions must not be defined without linguistic knowledge, because of the classification of contexts not appearing in the training data.

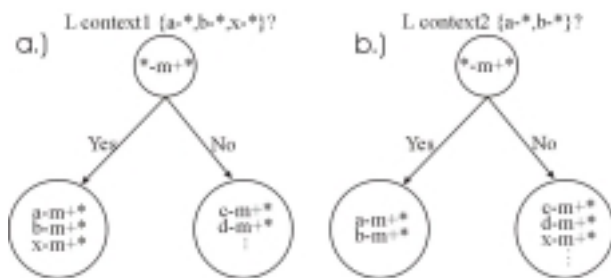


Figure 2. Effect of the order of questions on decision tree

5. Experimental methodology

The main scripts for training and testing acoustic models were implemented as Perl scripts invoking HTK. They were the outcome of the COST 249 project and intended to be used on the SpeechDat2 databases (Lindberg, 2000; Johansen, 2000) and are an extended version of the tutorial example in the HTK Book (Young et al., 2000). They can all be found on the Refrec homepage at

<http://www.telenor.no/fou/prosjekter/taletek/refrec/>

On this web page we can also find descriptions of standard tests and results of comparative tests done on many SpeechDat2 databases. We used hidden Markov models (HMM) having the 3-state left-right topology. We built triphone models and increased the number of Gaussian mixtures per state sequentially to 32.

5.1. The reference speech recognition systems

For building reference recognition systems we defined questions used in decision tree from groups of phones that were created as a part of the COST 249 project. For the English system we had 17 groups and for Slovenian 45 groups. During the training of acoustic models, data from labelled pronunciations of 800 speakers were used, while the data of the remaining 200 speakers was used as a test set.

The choice of good threshold values is important for the decision tree building process and requires some experimentation in practice. We therefore decided to experiment with the threshold set with the HTK RO command. This threshold determines how many training data each leaf in the decision tree must have. We built one Slovenian system with the threshold set to 100 and two English systems with thresholds set to 100 and 350, respectively (we named them sl-ref100, en-ref100 and en-ref350).

5.2. Speech recognition system with groups based on the SPE theory

In order to evaluate the effect of including the SPE theory into the decision tree building process we built five additional systems – three Slovenian and two English ones. For the model training we used the modified phonetic transcriptions as described in sec. 3.3 and 3.4. We automatically generated all natural phonetic groups from the distinctive feature compositions tables for both languages. From these groups, questions were generated that were used in the process of building decision trees for the two languages. Because of the modified phonetic transcriptions (less phones were used) and the modification of the broad classes, the number of leaves in the decision tree also changed and with that the distribution of the training data. In attempt to alter the amount of training data, we changed the threshold set with the HTK RO command for Slovenian systems from 100 to 267 and 350 and for English to 350 and 477. In this way we got five systems named sl-spe100, sl-spe267, sl-spe350, en-spe350 and en-spe477.

6. Speech recognition results

Six standard tests defined in the framework of the SpeechDat project (Johansen, 2000) were used on all reference and SPE based systems. These tests had the self-explanatory names: Yes/No test, Digits test, Connected Digits test, Application Words test, City Names test and Phonetic Rich Words test. In all tests but one (Connected Digits), each spoken test utterance consists of only one word. Therefore the word error rate (WER) is equal to the sentence error rate (SER) in these cases. Best results of tests done on all systems are given in table 6 and 7.

From these tables it can be observed that the SPE based systems performed either better or at least as good as the reference systems for both languages. The only

exception was the Application Words test on the Slovenian systems. We should also take into consideration that Yes/No, Digits and Connected Digits tests only applied to a small part of the decision tree. Specifically, the vocabulary in these tests is very limited and only a small number of triphones are therefore used.

	sl-ref100	sl-spe100	sl-spe267	sl-spe350
Yes/no	0,63	0,63	0,63	0,63
Digits	3,85	3,85	3,85	3,30
Con. Digits	4,12	3,91	3,95	3,98
App. Words	3,20	3,38	3,74	3,38
City Names	7,65	8,16	7,14	7,14
Ph. R. Words	17,62	17,36	15,93	15,51

	en-ref100	en-ref350	en-spe350	en-spe477
Yes/no	0,00	0,00	0,00	0,00
Digits	3,98	3,98	2,84	2,84
Con. Digits	5,42	5,51	4,22	4,33
App. Words	3,53	3,72	3,72	3,53
City Names	6,21	6,21	7,91	6,21
Ph. R. Words	36,83	35,01	32,68	31,56

Table 6: Lowest WER achieved by the Slovenian and English speech recognition systems in all six tests

	sl-ref100	sl-spe100	sl-spe267	sl-spe350
Con. Digits	15,75	14,56	14,56	14,32

	en-ref100	en-ref350	en-spe350	en-spe477
Con. Digits	30,72	30,92	24,50	25,10

Table 7: Lowest SER achieved by the Slovenian and English speech recognition systems in Connected Digits test

Without doubt, the most reliable evaluation of the SPE based concept can be taken from the Phonetic Rich Words test, employing the largest vocabulary (1491 words for Slovenian and 3043 for English) and more than 710 utterances. This test involves a very big part of the decision tree. This test also gave us the biggest decrease of the WER when comparing the SPE based concepts with the reference systems. The results achieved on the English systems had even bigger impact on the WER. The difference in WER of the best reference system and the best SPE based system is for Slovenian 1,85% and for the English 3,45%. Also the SER achieved with the SPE based systems in the Connected Digits test is better than the one achieved with the reference systems. The impact is again much bigger for English.

7. Conclusions

Within bounds of our experimental set-up we observed an advantage to include the SPE theory as an expert linguistic knowledge into the speech recognition systems. In general we got better results with the SPE-based processing for the English systems than for Slovenian ones. Several possible reasons can be referenced for such behaviour. One is probably the definition of phone groups

for the reference systems. There were 45 phone groups defined in the Slovenian reference system while only 17 in the corresponding English one. Therefore, the increase in the number of natural groups resulting from the inclusion of the SPE theory had bigger impact on the English systems than on the Slovenian ones. Another possible reason is the presence of noise. Pronunciations in the Slovenian database were recorded in much higher presence of noise than the English ones. This could potentially have reduced the distinctive ability of some of the features used in the SPE theory.

One possible reason for achieving much better WER for the Phonetic Rich Words test and SER for the Connected Digits test with the English SPE based systems could be the fact that the English reference systems had much bigger error rates than the Slovenian ones. The lowest WER in the Phonetic Rich Words test achieved by the Slovenian reference system was 17,62% whereas it in case of English was 36,83%. The same was observed for the SER in the Connected Digits test (English reference system: 30,72%, Slovenian reference system: 15,75%).

From our experiments, we also concluded that groups of phones should never include actual "unnatural groups" and leave it to the decision tree building process to disregard them in favour of the more natural groups. That would present no significant problem to the classification of triphones that do appear in training data but would lead to the incorrect classification of triphones with contexts that do not appear in the training set.

Based on the experimental evidence we have shown that the creation of the natural groups of phonemes by the SPE theory could effectively be used in defining phone groups for the multilingual speech recognition system including multilingual triphone Markov models. When porting the HLT technology to a new target language, this provides us a promising alternative to the more widespread approach of using the union of phone group definitions from all languages (Zgank et al., 2001).

8. References

- Brøndsted, T., 1998. A SPE based Distinctive Feature Composition of the CMU Label Set in the TIMIT Database. *Technical Report IR 98-1001*, Center for PersonKommunikation, Institute of Electronic Systems, Aalborg University
- Chomsky, Noam, and Halle, Morris, 1968. *The Sound Pattern of English*. Harper & Row, Publishers New York, Evanston, and London.
- Johansen, F.T., N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, 2000. The COST 249 SpeechDat multilingual reference recogniser. *Paper for XL-DB*.
- Lindberg, B., F.T. Johansen, N. Warakagoda, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, 2000. A Noise Robust Multilingual Reference Recognizer Based on SpeechDat(II). *In Proc. ICSLP, International Conference on Spoken Language Processing*, Beijing.
- Odell, J.J., 1995. *The Use of Context in Large Vocabulary Speech Recognition*. Dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy. Queens' College.
- Šuštaršič, R., S. Komar, and B. Petek, 1999. *Illustrations of the IPA: Slovene*. Handbook of the International Phonetic Association: A Guide to the Use of the

- International Phonetic Alphabet. Cambridge University Press, 135-139.
- Toporišič, Jože, 2000. *Slovenska slovnica*. Maribor: Založba Obzorja.
- Žgank, A., B. Imperl, F.T. Johansen, Z. Kačič, and B. Horvat. 2001. Crosslingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering. *In Proc. EUROSPEECH, European Conference on Speech Communication and Technology*. Aalborg.
- Young, Steve, Kershaw, Dan, Odell, Julian, Ollason, Dave, Valtchev, Vatcho, and Woodland, Phil. 2000. *The HTK Book (for HTK Version 3.0)*. Cambridge: Entropic Cambridge Research Laboratory.