Speech and Image Understanding, Proceedings of 3rd Slovenian-German and 2nd SDRV Workshop, April 24-26, 1996, Ljubljana, Slovenia, pp. 87–96.

Slovenian Word Recognition

I. Ipšić, F. Mihelič, N. Pavešić

Faculty of Electrical Engineering
Laboratory for Artificial Perception
Tržaška c. 25, 61000 Ljubljana, Slovenia
E-mail:ivoi@fe.uni-lj.si

E. Nöth

Universität Erlangen–Nürnberg
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstr. 3, 91058 Erlangen, Germany
E-mail:noeth@informatik.uni-erlangen.de

Abstract

In this paper we describe the architecture of a Slovenian continuous speech recognizer. The procedures used within the development of the speech recognizer and the speech recognition experiments are described. For training and testing a Slovenian speech database is used which contains read air flight information retrieval sentences.

1 Introduction

This paper describes a Slovenian continuous speech recognizer. The work on the development of the recognizer is performed within the joint project in multilingual speech recognition and understanding *Spoken Queries in European Languages* (SQEL-Copernicus -1634). The aim of the joint project is to built a system capable, of having a dialogue in one of the four European languages German, Slovenian, Czech and Slovak about a task oriented topic. Such a system has to be able to handle spontaneous speech, and to provide the user with correct information. The information system being developed for Slovenian speech will be used for air flight information retrieval. The system has to answer questions about air flight connections and their time and date. The architecture of the Slovenian information retrieval system is based on the German dialog system, developed within the European Esprit project Sundial [Pec91]. The German demonstrator was further developed at University of Erlangen after the end of Sundial, and it can perform human machine continuously spoken dialogues via telephone. It can answer questions

about train connections. The architecture of the system consists of three main modules: the word recognition module, the linguistic analysis module and the dialog manager [EKN+93].

In this paper we describe the development of the word recognition module for Slovenian continuous speech. The development of the Slovenian continuous speech recognizer can be grouped into the following procedures:

1. feature vector extraction,
2. vector quantization and feature vector transformation,
3. acoustic modelling,
4. language modelling and
5. word recognition.

In the following sections these procedures and their implementation are described. First we present an overview of implemented procedures and in section 8 we describe the Slovenian word recognition experiments and results.

2 Feature Vector Extraction

For Slovenian word recognition we have tested several feature representations. According to several speech recognition systems [NP88, LHH+89] we have evaluated cepstrum features [DM80] for Slovenian speech recognition. In [Mih91, MIDP92] we chose MEL-cepstrum and LPC-cepstrum coefficients as the most suitable feature representations for Slovenian speech. Several experiments have proved that adding dynamic features to the feature vectors increases the recognition accuracy. We have performed word recognition using different sets of feature vectors, which consist of the speech frame energy, cepstrum coefficients and their first order and second order derivatives. The number of cepstral coefficients is reduced from 128 to 11 by integrating spectral bands according to the Mel scale. Cepstral coefficients calculated on the basis of Fourier spectrum showed slightly better results then coefficients based on LPC spectrum. Best results were achieved when the feature vectors consist of frame energy, MEL-cepstrum coefficients and their first and second order regression coefficients.

3 Vector Quantization and Feature Vector Transformation

The purpose of vector quantization is to reduce the amount of information and to represent the speech signal in a different structure, more suitable for recognition. Using vector quantization techniques feature vectors are grouped into clusters. The clusters can be represented by a prototype vector, a density function or by a mixture density function. In the training phase first all feature vectors are clustered into 64 classes. Then using the quantized feature vectors and a phonetic lexicon the training sentences are labeled with 31 phone symbols. For the labeling procedure every phone is represented by a simple hidden Markov model. Using the phonetic

lexicon the training sentences are represented with phone models and 10 Baum-Welch training iterations are used to estimate the HMM parameters. Using the labeled training sentences, feature vectors belonging to same phones are grouped into phone clusters. In a second quantization step every phone class cluster is divided into 5 subclusters. For this we use the K-means algorithm, and each class is represented by a multivariate Gaussian probability density function. So we get 31x5=155 speech frame clusters to quantize the speech feature vectors.

To increase the separability of feature vector clusters or classes we have tested linear discriminant methods. Experiments have shown that the use of linear discriminant analysis increases the recognition accuracy for Slovenian phonemes. Therefore we have also evaluated the discriminant analysis method on continuous word recognition. We propose a transformation procedure, where the classes for the transformation are defined as phone units. For the transformation procedure the training vectors were clustered into phone classes. The criterion used for defining the linear transformation is the maximization of phone classes separability [DH73, KY73]. Recognition results using different dimension reductions are shown in section 8.

4 Acoustic Modelling

To model the different speech events like phonemes, context dependent phonemes, words and sentences we use the ISADORA system [ST94]. As basic speech units we choose phone components. Using them Slovenian phonemes can be defined. HMMs of phonemes are built by sequential concatenation of phone component models. In the same way word models are built by concatenation of phoneme models. To model the phonemes and words the system uses basic speech units models, which are represented by a linear HMM, where each state is connected to itself and to its successors.

To model the different coarticulation of phonemes we use context dependent phone models, the polyphones [STKN94]. A polyphone consists of a phone with arbitrary length of left and right context of phones, and it is determined from the training database. The only criterion to form a polyphone is the minimal number of its occurrences in the words of the training sentences. Polyphones, which consist of a large context, and have a high number of occurrences describe the speech signal in a better way than models generated by simple concatenation of monophone models. The phonetic context of a polyphone unit may also span over entire words if their occurrence in the training material is high enough. Different speech units like polyphones or words are built by concatenation or parallelization of basic speech units, which are assumed to be acoustically stationary events. In such a way the hierarchical structure of speech is transformed into a network of polyphone HMMs.

To train the polyphone models we have to estimate the transition probabilities and the output density parameters. The goal of the training procedure is to maximize the probability of the polyphone model given the observation sequence. To evaluate the parameters of the model we used an iterative training procedure (APIS)[STKN94] which is a modified version of the Baum-Welch algorithm. The idea of the training procedure is that the polyphones and their more general models cover the same parts of the speech signal.

5 Language Modelling

The language model determines the a priori probability of a word sequence. The probability of the word sequence can be determined from a product of conditional probabilities, where we need for each word from the lexicon a probability that it follows any possible word sequence [STKN94]. Since it is impossible to determine conditional probabilities of various lengths, the word sequence probability has to be approximated with conditional probabilities, where each word depends only on some previous words. If we use $N-1$ words to approximate the conditional word probabilities we call it a N-gram language model. Conditional word probabilities can be estimated by relative frequencies of different word sequences from the training corpus. Since many of the possible word sequences do not occur in the training corpora it is necessary to solve the problem of missing word sequences. One of the possible approaches is to interpolate the probabilities of higher order language models with the lower order N-gram probabilities [JM80] i. e. tri-gram with bi-gram and uni-gram probabilities. Another approach to the problem of missing word sequences is the use of word categories [KDM90]. Words from the recognition vocabulary are assigned to word categories, and so the number of parameters of the language model can be drastically reduced.

We have defined 127 word categories to classify the words from the Slovenian recognition vocabulary. The categories group words with same grammatical and semantical characteristics [Gro94]. The bigrams were trained on 10000 sentences comprising 800 different vocabulary words.

6 Word Recognition

For word recognition word models are constructed by concatenation of phoneme models, which again are obtained by concatenating the polyphone models according to the pronunciation lexicon. All word models are concatenated in parallel and form a single Hidden Markov Model, which is represented by a huge network of nodes. The analysis of an unknown observation sequence is performed by the Viterbi algorithm, producing the maximum a posteriori state sequence of the model with respect to the observed input vectors. Knowing the state sequence of the HMM we can decode the input sequence and transform it into a string of words. Because of the large number of states which have to be considered when computing the Viterbi alignment, a state pruning technique has to be used to reduce the size of the search space. We use the Viterbi beam–search technique which expands the search only to states which probability falls within a specified beam. The probability of reaching a state in the search procedure cannot fall short of the maximum probability by more than a predefined ratio. During the forward search in the HMM N best word sequences are generated using acoustic models and a bigram language model. The N best word sequence hypotheses are then reordered using the sentence probability of higher order language models [Ste91, Kuh94].

Figure 1 illustrates the recognition procedures. First every 10 ms speech vectors are computed and then transformed. In the next step vectors are quantized into 155 clusters, which are represented by multivariate Gaussian distributions. Quantized speech vectors are then passed

to the hidden Markov model of the words from the vocabulary. N best word sequence are generated and then re scored using a trigram language model.

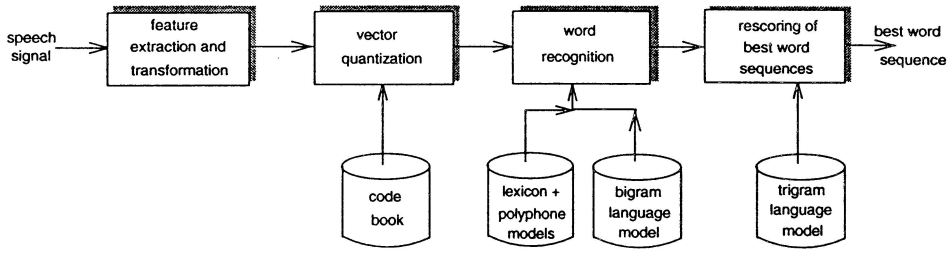

Figure 1: Architecture of the Slovenian word recognizer.

7 The Slovenian Database

The application domain of the Slovenian speech database covers flight information retrieval dialogues. A corpus of sentences was selected from transcribed recordings made at the booking center of the Adria Airways company in Ljubljana. The selected sentences were grouped in four categories: introductory and concluding parts of the dialogue, central part of the dialogue concerning the selected information domain, questions that would be redirected to another address and utterances consisting of words determining time and date. The database consists of approximately 5000 sentences with about 800 different words. The sentences were reread by 50 speakers (25 female and 25 male), and were recorded on a HP Workstation with 16kHz sampling rate. Every speaker has spoken about 170 sentences. The recordings have been simultaneously performed over the telephone line, so additionally 7 hours of telephone quality speech were obtained [MDG+94].

8 Recognition Results

Several experiments using the previously mentioned database were performed: recognition tests using different sets of feature vectors and tests on transformed feature vectors, recognition tests using different acoustic models and speaker dependent and speaker independent recognition using the language model.

In the tables we show the word accuracy (WA), the sentence accuracy (SA), and error rates for substituted words (S), inserted words (I) and deleted words (D). Word accuracy is computed

using the formula:

$$WA = 100(1 - \frac{S + D + I}{N})\%, \quad (1)$$

where S, D, I are the substituted, deleted and inserted words, and N is the number of spoken words. The sentence accuracy SA gives the percentage of correctly recognized whole sentences.

In Tables 1-4 we compared different feature sets and the feature vector transformation procedure. For training we use polyphone models of maximum context of 2. For recognition no language model is used. Speaker dependent results are given for microphone speech, while the speaker independent results are given for telephone quality speech. The speech signal is sampled with 16000 kHz and cepstrum features and their derivatives are computed every 10 ms. The cepstrum derivatives were approximated using first and second order regression coefficients [Fur86, AH91]. The regression coefficients were computed over 5 neighboring speech frames. Table 1 shows speaker dependent word recognition results obtained using different feature vector representations. The recognizer was trained on one speaker, and the same set of sentences was used for training and testing. The first feature set contains the frame energy and 11 cepstral coefficients. The second set was extended with 12 first order regression features. Table 2 shows speaker independent recognition results of telephone speech using different feature representations. The recognizer was trained on speech from 10 speakers (1700 sentences with 11500 words), and another set of sentences spoken by 10 speakers (1700 sentences with 11500 words) was used for testing. Best results are obtained when the feature vector, which contains the frame energy and 11 cepstral coefficients is extended with 12 first and 12 second order regression coefficients.

feature	dimension	$WA(\%)$	$SA(\%)$	$S(\%)$	$D(\%)$	$I(\%)$
LPC	12	86.3	51.1	5.6	8.0	0.0
MEL	12	86.6	51.7	7.9	5.3	0.1
LPC + slopes	24	94.5	87.5	4.2	1.2	0.0
MEL + slopes	24	96.9	93.0	1.8	1.2	0.0

Table 1: Speaker dependent recognition results.

feature	dimension	$WA(\%)$	$SA(\%)$	$S(\%)$	$D(\%)$	$I(\%)$
MEL + slopes	24	58.9	25.3	26.9	12.8	1.5
MEL + slopes	36	63.6	27.1	23.4	12.2	0.9

Table 2: Speaker independent recognition results.

Tables 3 and 4 show the influence of the vector transformation technique on Slovenian word recognition. When reducing the feature vector dimension from 24 to 12 features word accuracy does not significantly change, while the sentence accuracy increases by 30% in comparison with recognition results obtained with a 12 dimensional feature vector (Table 1). Speaker independent recognition results show only a small increase of word and sentence accuracy when comparing recognition results obtained by reducing the feature vector from 36 to 24 features

with recognition results of untransformed features (Table 2). When reducing the feature vector from 36 to 20 features the recognition results are still slightly better in comparison with results obtained with 24 dimensional untransformed feature vectors.

feature	dimension reduction	$WA(\%)$	$SA(\%)$	$S(\%)$	$D(\%)$	$I(\%)$
LPC	24– >20	93.3	88.8	5.4	1.2	0.0
MEL	24– >20	94.5	88.8	4.8	0.6	0.0
LPC	24– >12	87.9	79.1	8.4	3.6	0.0
MEL	24– >12	88.8	58.1	5.3	5.8	0.0

Table 3: Speaker dependent recognition results using feature vector transformations.

feature	dimension reduction	$WA(\%)$	$SA(\%)$	$S(\%)$	$D(\%)$	$I(\%)$
MEL	36– >24	59.9	28.8	24.7	14.9	0.3
MEL	36– >20	59.8	26.2	26.7	12.1	1.2

Table 4: Speaker independent recognition results using feature vector transformations.

Table 5 shows recognition results using different acoustic models. The recognizer is trained on microphone speech spoken by 10 speakers. Recognition results are given for speaker independent recognition without using a language model. Best recognition results are obtained using the polyphone models. No significant improvement is achieved when the context of the polyphones is larger than two left and two right phones. So we use polyphones with maximum context of two phones.

model	$WA(\%)$	$SA(\%)$	$S(\%)$	$D(\%)$	$I(\%)$
monophones	37.5	13.3	45.4	16.0	1.0
monophones+ frequent word models	66.2	29.4	19.8	13.1	0.7
biphones+triphones+ frequent word models	71.3	35.6	19.3	6.3	2.9
polyphones+ frequent word models	75.7	37.0	14.1	9.5	0.5

Table 5: Recognition results using different acoustic models.

Table 6 presents speaker independent word recognition results when using polyphone acoustic models and a language model. In the search procedure a word bigram with perplexity 10 is used. The best sentence hypotheses are reordered using a trigram language model. In table 6 we show average recognition results for female and male speakers obtained from microphone as well as from telephone signals.

speaker	signal	$WA(\%)$	$SA(\%)$	$S(\%)$	$D(\%)$	$I(\%)$
female	mic	83.9	50.5	9.5	3.6	2.7
male	mic	84.1	56.6	9.2	5.1	1.5
female	tel	75.2	43.3	15.2	6.2	3.2
male	tel	77.5	45.0	14.4	5.2	2.8

Table 6: Speaker independent recognition results using polyphone models and language models.

9 Conclusion

We presented procedures used for the development of a Slovenian speech recognizer for a specified communication domain. The recognizer has been developed within the SQEL project, where a information retrieval system for air flight information is being developed.

For word recognition we have tested several feature representations. Best recognition results were achieved when the MEL-cepstrum feature vectors are extended with first and second order regression coefficients. We have shown that the recognition accuracy can be increased when using a linear feature vector transformation procedure.

For statistical subword and word modelling we use the ISADORA system. To model the words, we want to use in the speech recognizer, polyphone models are generated and trained, and then used to build word models. Experiments have shown that the context of 2 of the polyphone models guarantees high recognition accuracy.

Using stochastic language models the recognition accuracy is significantly increased. A word bigram language model with perplexity 10 is used in the search procedure. Additionally best word sequences are reordered using a word trigram language model. Also a category based language model is used but no significant recognition improvement has been achieved, since the perplexity of the word language model is low.

The recognizer has been evaluated on microphone signals as well as on telephone speech. The recognition results obtained for telephone quality speech are similar to that reported in [EKN+93]. Thus the word recognition system would be appropriate for the Slovenian information system, which could enable a dialogue with a user over the telephone line.

References

[AH91] T.H. Applebaum and B.A. Hanson. Tradeoffs in the Design of Regression Features for Word Recognition. In *Proc. European Conf. on Speech Technology*, volume 3, pages 1203–1206, 1991.

[DH73] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[DM80] S.B. Davis and P. Mermelstein. Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[EKN+93] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E.G. Schukat-Talamazzini. A Spoken Dialogue System for German Intercity Train Timetable Inquiries. In *Proc. European Conf. on Speech Technology*, pages 1871–1874, Berlin, 1993.

[Fur86] S. Furui. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 34(1):52–59, 1986.

[Gro94] J. Gros. Postavljanje hipotez o stavkih pri razpoznavanju vezanega slovenskega govora. M.Sc. Thesis, Univerza v Ljubljani – Fakulteta za elektrotehniko in računalništvo, Ljubljana, 1994.

[JM80] F. Jelinek and R.L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice*, pages 381–397. North Holland, 1980.

[KDM90] R. Kuhn and R. De Mori. A Cache-Based Natural Language Model for Speech Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(6):570–583, 1990.

[Kuh94] T. Kuhn. Die Erkennungsphase in einem Dialogsystem. Dissertation IMMD5 (Mustererkennung), Universität Erlangen, 1994.

[KY73] J. Kittler and P.C. Young. A New Approach to Feature Selection Based on the Karhunen-Loeve Expansion. *Pattern Recognition*, 5:335–352, 1973.

[LHH+89] K.-F. Lee, H.-W. Hon, M.-Y. Hwang, S. Mahajan, and R. Reddy. The SPHINX Speech Recognition System. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 445–448, 1989.

[MDG+94] F. Mihelič, S. Dobrišek, J. Gros, I. Ipšić, P. Pepelnjak, and N. Pavešić. Development of an Continuous Speech Recognition System for Information Services. In B. Horvat and Z. Kačič, editors, *Modern Modes of Man-Machine Communication*, pages 17–1 – 17–20, Maribor, 1994.

[MIDP92] F. Mihelič, I. Ipšić, S. Dobrišek, and N. Pavešić. Feature Representations and Classification Procedures for Slovenian Phone Recognition. *Pattern Recognition Letters*, 12(12):879–891, 1992.

[Mih91] F. Mihelič. Akustično fonetična pretvorba slovenskega govora. Ph.D. Thesis, Univerza v Ljubljani – Fakulteta za elektrotehniko in računalništvo, Ljubljana, 1991.

[NP88] H. Ney and A. Paeseler. Phoneme-Based Continuous Speech Recognition Results for Different Language Models in the 1000-Word SPICOS System. *Speech Communication*, 7(4):367–374, 1988.

[Pec91] J. Peckham. Speech Understanding and Dialogue over the Telephone: an Overview of Progress in the SUNDIAL Project. In *Proc. European Conf. on Speech Technology*, volume 3, pages 1469–1472, 1991.

[ST94] E.G. Schukat-Talamazzini. *Automatische Spracherkennung*. Vieweg, Wiesbaden, 1995.

[Ste91] V. Steinbiss. A Search Organization for Large-Vocabulary Recognition Based on N-Best Decoding. In *Proc. European Conf. on Speech Technology*, volume 3, pages 1217–1220, 1991.

[STKN94] E.G. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialog Systems. In H. Niemann, R. de Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology*, pages 110–120. CRIM/FORWISS, Infix, 1994.