A 10 000-WORD CONTINUOUS-SPEECH RECOGNITION SYSTEM

V. Steinbiss, A. Noll, A. Paeseler, H. Ney, H. Bergmann,
C. Dugast, H.-H. Hamer, H. Piotrowski, H. Tomaschewski, A. Zielinski

Philips GmbH Forschungslaboratorium Hamburg
D-2000 Hamburg 54, West Germany

ABSTRACT

This paper reports on some results we obtained when increasing the recognition vocabulary size of our phoneme-based speaker-dependent continuous-speech recognizer from 1 000 to 10 000 words. The potential search space increased from 46 000 to 516 000 states without problems for the data-driven search. We focused our attention on two points: The performance of our phoneme models and the impact of the language model. The main results are:

(1) Increasing the recognition vocabulary by a factor of 10 (from a perplexity of 917 to 9686) increased the word error rate by a factor of 2 (from 21.8% to 43.1%). We tested phoneme models with both discrete probabilities and continuous mixture densities. The mixture density models performed better; moreover, they saved about half of the search costs.

(2) The language model was found to be very important for a larger vocabulary size. With a test set perplexity of 388 (i. e. a reduction by a factor of 25 compared to the case without bigram model) the error rate decreased by a factor of 2.4. In order to check how meaningful perplexity is for the prediction of the system's performance, we constructed a stochastic language model with a perplexity of 1000, the size of the vocabulary used in previous experiments, and got about the same error rate.

1. INTRODUCTION

The speaker-dependent continuous-speech recognizer described in [4,7] has until now been used with a (German) 917-word recognition vocabulary. Our goal was to investigate the effects of increasing the recognition vocabulary size from 1 000 to 10 000 words, namely on the performance of our phoneme models and on the data-driven search. Another point of our interest was the impact of language modelling on a very large vocabulary.

The recognition system is based on statistical principles and on Viterbi approximation - most likely state sequence - in both training and recognition. The main features of the system and its environment can be summarized as follows:

Training and test data [9]:

- Recorded in an office environment, close-talking microphone
- Read speech
- 2 × 100 phonetically balanced training sentences comprising 2 × 524 words (a total of 7 minutes of speech)
- 200 test sentences (database queries) comprising 1391 words
- Small overlap (43 words) of training and testing vocabularies.

Preprocessing [4]:

- Sampling rate 16 kHz
- 30 cepstrally smoothed spectral intensities in logarithmic units, normalized with respect to average intensity, plus intensity
- Additionally, first and second temporal differences of these.

Acoustic-phonetics [3]:

- Standard pronunciation dictionary
- 44 context-independent phonemes
- Variants of Hidden Markov Models (HMMs) with both discrete probabilities and continuous mixture densities (for details, cf. below).

Language Model [6]:

- Recognition vocabularies comprising 917 or 9686 words
- Either no language model constraints or stochastic bigram language models.

Search [2]:

- Data-driven one-pass dynamic programming search.

2. TEXT CORPORA AND VOCABULARIES

The system was trained on two sessions of 100 phonetically balanced German sentences which are known as Sotscheck sentences or Berlin sentences [8,9]. This training vocabulary is fairly different from the set of words spoken in the test sessions (the test vocabulary). The two recognition vocabularies taken into account are

- the 917 SPICOS words and
- a 9686-word vocabulary including the SPICOS words

(cf. Table 1). The second vocabulary was derived from a lexicon made up at the University of Bochum within the ESPRIT project "Linguistic Analysis of the European Languages" (project no. 291/860). This lexicon consists of words drawn from newspaper and ESPRIT texts and comprises, for each word, one single (standard) phonetic transcription and a morpho-syntactic labelling.

CH2847-2/90/0000-0057 $1.00 © 1990 IEEE

In informal experiments we had to observe that the merging of different pronunciation lexica must be done with care: Inconsistencies between training and testing vocabularies led to a significant deterioration of performance, as compared to a lexicon with a consistent phonetic transcription.

Table 1. Training vocabulary and the two test vocabularies used in the system.

Speech corpus	Sotscheck words	SPICOS vocabulary	10 000-word vocabulary
Function	Training	Recognition	Recognition
No. of words	341	917	9686
Average no. of phonemes per word	4.8	8.5	9.0
Overall no. of states	-	45 898	516 211

3. ACOUSTIC-PHONETIC MODELLING

General

In our phoneme-based recognition system, each word in the lexicon has one single phonetic transcription. Each of these context-independent phonemes is expanded into (e.g., three) phoneme segments in order to incorporate coarticulatory effects implicitly into the model. Each segment is then expanded into a sequence of states for a better temporal modelling. The phonemes are variants of Hidden Markov Models (HMMs) with fixed transition probabilities. We did Viterbi training using only 7 minutes of speech. Both the discrete approach and the mixture-density approach were tested within the same framework but differ in several details (e.g., the number of segments per phoneme) which are due to system optimizations.

Discrete Modelling

As compared with [5,3], our discrete approach has been improved in several directions. Like the mixture-density approach, its feature vectors cover besides the logarithmic spectral energies their first and second temporal differences. As this increase in dimensionality (in fact, a factor of 2) calls for a much bigger vector quantization (VQ) codebook and thus for much more training data, we worked with three different codebooks for the frequency pattern vector as well as for the first and second order differences, thus assuming that these three parts are statistically independent. Typical codebook sizes for the three feature vector subspaces are 512 (for the logarithmic spectral energies), 256 (first temporal differences), and 256 (second temporal differences).

The training is initialized with a (triple) VQ codebook derived from all of the four (two male, two female) speakers; in tests this turned out to be better than taking a (triple) codebook derived from each single speaker. In each iteration of the Viterbi training, the probability density functions (pdfs) were reestimated (moderately smoothed); during the first iterations, we also reestimated the prototype vectors of the codebook using the assignment of the observed vectors based on the Viterbi alignment.

Mixture Densities

The parametric mixture density models are described in [3]. For each HMM state, the probability density function of emitting a vector $\mathbf{x}$ is

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k \, p_k(\mathbf{x})$$

where k is the index of the mixture components, with weights c_k and parametric probability density functions $p_k(\mathbf{x})$. In this system, the $p_k(\mathbf{x})$ are Laplacian distributions with a fixed vector of absolute deviations.

4. THE LANGUAGE MODELS

Our language model is a stochastic bigram model [6] based on word categories, i. e. the conditional probability

$$p(w_n|w_1 \ldots w_{n-1})$$

of observing the word w_n after a word sequence $w_1, \ldots, w_{n-1}$ is assumed to be

$$p(w_n|w_{n-1}) \approx p(w_n|C_n) \cdot p(C_n|C_{n-1})$$

where C_i denotes w_i's word category and where the N_i words in the category C_i are assumed to be equiprobable: $p(w_i|C_i) = 1/N_i$. Two language models were constructed by using two different training sets:

a) 207 independently collected sentences (comprising 1886 words) of the SPICOS test sentence type (database queries), giving a test set perplexity of 388.

b) The same sentences plus further 100 000 words of newspaper text (weighted twice). The resulting test set perplexity of 1003 is close to the SPICOS vocabulary size of 917.

87% of the category bigrams seen in the test sentences are covered by the corpus in a) and 95% by the corpus in b). In both cases we used a set of morpho-syntactic task-independent word categories (cf. Table 2) made up within the ESPRIT project 291/860. The choice of these categories is motivated by two reasons: On the one hand it is necessary to have a detailed representation of so-called function words, like determiners and prepositions, that occur very frequently in every language. On the other hand German words can have several inflectional derivations containing e. g. information about case and gender. This is why the same word form can be a member of several categories.

We use 212 of the total number of 355 categories to model the closed word classes, including particles, conjunctions, determiners, prepositions, pronouns and punctuation marks. These categories contain less than 30 words each, in most of the cases only one or two words.

The remaining 143 categories model the open word classes, including common nouns, proper nouns, adverbs, adjectives, verbs, numbers and abbreviations. These categories contain from one to more than 1000 words and subdivide each part of speech into several morphological classes, e. g. according to all possible combinations of case, gender and number in the case of common nouns.

Table 2 shows, for different parts of speech, their number of categories and words. Due to multiple countings, the sum of words exceeds the vocabulary size of 9415. This figure (9415 instead of 9686), which also occurs in Table 6, is due to the fact that the labelling of words not covered by the training texts had not been completed when the experiments were done.

Table 2. Number of categories and of words per part of speech.

Part of speech	No. of categories	No. of words
Adjective	33	2 045
Common noun	38	4 653
Proper noun	16	1 081
Full verb	18	2 138
Auxiliary verb	30	77
Particle / adverb	8	301
Determiner	36	42
Pronoun	139	375
Conjunction	15	84
Preposition	6	125
Miscellaneous	15	82
End-of-sentence mark	1	1
Total:	355	11 004

5. EXPERIMENTAL RESULTS

Experimental tests were made for four speakers on one test session each. Apart from minor changes due to the enlarged memory requirements and the new pronunciation lexicon, there was no special adaptation of the system to the new task. The data-driven search was able to handle the larger search space, which increased from 46 000 to 516 000 states, without any problems. For both no language-model constraints and the bigram model the actual size of the search space was reduced to about 10% to 20% of the potential search space by using a pruning technique.

Table 3. Recognition results (Del = deletions, Ins = insertions, Sum = word error rate) with discrete models and without language model restrictions for two recognition vocabularies:
a) vocabulary of 917 words,
b) vocabulary of 9686 words.

Speaker	Search	Paths lost	Del	Ins	Sum
a)					
M-03	15 000	2	2.4%	1.4%	15.6%
M-10	15 000	4	5.3%	2.7%	31.2%
F-01	18 000	0	7.0%	1.4%	32.9%
F-10	12 000	4	3.4%	1.4%	22.4%
average:					25.5%
b)					
M-03	150 000	0	6.2%	3.8%	42.6%
M-10	156 000	0	17.2%	2.7%	74.2%
F-01	190 000	0	11.3%	4.8%	64.5%
F-10	152 000	0	10.1%	2.7%	48.8%
average:					57.5%

Recognition results are shown in Tables 3 to 5 (summarized in Table 6). "Search" denotes the average number of states per centisecond; this figure is followed by the number of sentences where the optimal path through the spoken word sequence is lost due to pruning. The test set consists of 200 sentences comprising 1391 words per speaker. The word error rate ("Sum") covers deletions ("Del"), insertions ("Ins") and substitutions of words.

Table 4. Recognition results (Del = deletions, Ins = insertions, Sum = word error rate) with mixture densities and without language model restrictions for two recognition vocabularies:
a) vocabulary of 917 words,
b) vocabulary of 9686 words.

Speaker	Search	Paths lost	Del	Ins	Sum
a)					
M-03	6 365	4	2.7%	1.4%	15.7%
M-10	7 127	0	4.4%	3.2%	25.7%
F-01	9 274	0	3.5%	2.4%	28.3%
F-10	7 560	4	2.4%	1.3%	17.6%
average:					21.8%
b)					
M-03	89 964	0	4.2%	4.1%	35.0%
M-10	82 188	0	5.7%	9.3%	43.3%
F-01	106 839	0	6.8%	6.1%	54.6%
F-10	105 637	0	4.6%	4.3%	39.5%
average:					43.1%

Table 5. Recognition results (Del = deletions, Ins = insertions, Sum = word error rate) with mixture densities and different stochastic bigram language models on the big recognition vocabulary
a) perplexity = 388, language model trained on SPICOS-like sentences,
b) perplexity = 1003, language model trained on SPICOS-like sentences and newspaper text.

Speaker	Search	Paths lost	Del	Ins	Sum
a)					
M-03	44 171	4	1.4%	0.3%	10.6%
M-10	37 398	5	3.4%	1.9%	22.4%
F-01	58 590	4	3.2%	0.9%	23.4%
F-10	54 969	4	1.9%	1.3%	14.2%
average:					17.6%
b)					
M-03	60 262	5	2.8%	0.8%	15.3%
M-10	74 904	1	3.9%	1.9%	24.2%
F-01	79 643	1	5.2%	1.7%	32.6%
F-10	73 360	2	3.7%	1.7%	22.0%
average:					23.5%

Table 6. Breakdown of all test results (averaged over four speakers).

Vocabulary size	Test set perplexity	Emission pdf type	Error rate
917	917	discr.	25.5%
917	917	cont.	21.8%
9 686	9 686	discr.	57.5%
9 686	9 686	cont.	43.1%
9 415	1 003	cont.	23.5%
9 415	388	cont.	17.6%

For the tests without language model, the increase in the recognition vocabulary size by a factor of 10 caused an increase by a factor of 2 in the average error rate (Tables 3 and 4). The higher number of errors is mainly due to the additional short words in the big lexicon. Table 7 shows recognition examples. For the mixture density models (Table 4), the average word error rate increased from 21.8% to 43.1%, compared to an increase from 25.5% to 57.5% for the discrete models (Table 3). Thus, the mixture densities performed better than the discrete models.

Table 7. Recognition examples for different language models.
a) Spoken sentence. b) - e) Recognized sentences: b) and c) without language model, d) and e) with bigram language model.
Perplexities: b) 917, c) 9686, d) 1003, e) 388.

a) *von wann ist das letzte Rundschreiben*
b) *von waren es das letzte Rundschreiben*
c) *form waren es Tass letzte Rundschreiben*
d) *von wann ist das letzte Wunsch waren*
e) *von wann ist das letzte Rundschreiben*

a) *wieviele Anträge an das BMFT gibt es*
b) *wieviele Anträge eines BMFT gib des*
c) *wie vieler Anträge anders BMFT gib des*
d) *die vieler Anträge an das BMFT gibt es*
e) *wieviele Anträge an das BMFT gibt es*

The experiments also underline the importance of a language model for a large-vocabulary recognition task. Using a bigram model with a perplexity of 388 which had been trained on (different) sentences of the same type (database queries), the error rate could be reduced by a factor of 2.4 (down from 43.1% to 17.6%, cf. Table 5a). Note that this significant gain in performance was bought at the price of a decrease in perplexity by a factor of 25 (from 9686 to 388).

It is known [1] that a first approximation of the difficulty of a recognition task is the test-set perplexity. This was confirmed by tests with a bigram model with about the same perplexity (1003) as the smaller recognition task (917), where we indeed found error rates within the same range (Tables 4b and 3a).

REFERENCES

[1] F. Jelinek: "The Development of an Experimental Discrete Dictation Recognizer", Proc. of the IEEE, Vol. 73, No. 11, pp. 1616-1624, Nov. 1985.

[2] H. Ney, D. Mergel, A. Noll, A. Paeseler: "A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition", Proc. 1987 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Dallas, Texas, pp.20.10.1-4, April 1987.

[3] H. Ney, A. Noll: "Phoneme Modelling Using Continuous Mixture Densities", Proc. 1988 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, New York, pp. 437-440, April 1988.

[4] H. Ney, A. Paeseler: "Phoneme-Based Continuous Speech Recognition Results For Different Language Models in the 1000-Word SPICOS System", Speech Communication 7, pp. 367-373, 1988.

[5] A. Noll, H. Ney: "Training of Phoneme Models in a Sentence Recognition System", Proc. 1987 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Dallas, Texas, pp. 29.6.1-4, April 1987.

[6] A. Paeseler, H. Ney: "Continuous Speech Recognition Using a Stochastic Language Model", Proc. 1989 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Glasgow, UK, pp. 719-722, May 1989.

[7] A. Paeseler, V. Steinbiss, A. Noll: "Phoneme-based Continuous-Speech Recognition in the SPICOS-II System", to be published in IT, Series "Computer, Systeme, Anwendungen", R. Oldenbourg Verlag, 1989.

[8] J. Sotscheck: "Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die deutsche Sprache", Proc. DAGA '84, Deutsche Arbeitsgemeinschaft für Akustik, Darmstadt, West Germany, 4 p., March 1984.

[9] V. Steinbiss, H.-H. Hamer, D. Mergel, H. Ney, A. Noll, A. Paeseler, H. Piotrowski, H. Tomaschewski: "The Speech Database Used in SPICOS", Proc. of the ESCA Workshop on 'Speech Input/Output Assessment and Speech Databases', pp. 2.7.1-4, Noordwijkerhout, the Netherlands, Sept. 1989.