PHONEME-BASED RECOGNITION FOR THE NORWEGIAN SPEECHDAT(II) DATABASE

Finn Tore Johansen

Telenor Research and Development, N-2007 Kjeller, Norway
finn.johansen@fou.telenor.no

ABSTRACT

This paper presents results from a number of flexible vocabulary recognition experiments on the Norwegian SpeechDat(II) database. A common phoneme-based recogniser design procedure is tested on five different tasks, and for five different training sets. Results verify that reasonably accurate recognisers can be built with the database, using standard HMM techniques. They also quantify the importance of training set selection for small and medium vocabulary tasks.

1. INTRODUCTION

The SpeechDat(II) project (LE2-4001) [1] is recording 28 speech databases in 21 different European languages. The content of these databases has been specified to fit the current and near future needs for telephone service development and testing. They contain small vocabularies, such as digits, letters and command words, as well as phonetically rich isolated words, names and sentences intended for medium and large vocabulary speech recognition applications. All recordings are made over the telephone network, with a controlled distribution over age, sex and dialect.

The existence of such large, multilingual speech corpora, designed from common specifications, presents exciting opportunities for speech recognition research and development, as shown by previous work on the similar, but older Polyphone (e.g. [2, 3]) and SpeechDat(M) (e.g. [4]) databases.

This paper is an extension of the work reported in [5] on the Norwegian fixed network database (FDB) [6]. The basic approach is to train a set of phone-based acoustic models on the entire database, and test these models on different tasks, with different vocabulary sizes and complexities.

In [5], multiple tests on four different tasks were used to conlude that decision tree clustered triphones outperformed context independent models and that flat start training did not always perform as well as models initialised from a manually segmented database (EUROM). Differences between vocabulary independent and vocabulary dependent training, as well as two different feature sets, were also quantified.

In this paper, we report a new set of results on train-set dependency. The decision tree clustered triphone modelling technique, along with EUROM-initialised models and the best feature set from [5] is retained. Variance floors have however been added to avoid problems with high number of mixture components, and grammar scaling is now used in the connected digit test. A fifth, medium-size vocabulary test with city name recognition is also included, along with a phonetically rich, isolated word only training condition.

TST	CCD	TSZ	VSZ	SCM	ACN
01	I1	173	12	SENTENCE MATCH	173
02	Q1-2	376	2	SENTENCE MATCH	376
03	A1-6	1174	30	SENTENCE MATCH	1174
04	B1,C1,C4	576	12	WORD MATCH	4601
05	O2-4	561	432	WORD MATCH	602

Table 1: Five tests for the Norwegian FDB. See text for explanation of mnemonics.

In Section 2 the five tests are summarised. Sections 3 and 4 describe the recogniser design procedure and different sub-corpora used for training. Results and conclusions are given in Section 5 and 6.

2. TEST DEFINITIONS

All SpeechDat databases for fixed and mobile networks contain an official list of recording sessions for testing [7]. In the Norwegian 1016 speaker FDB this list contains 200 sessions. A common test design *format* is also defined in [7] to facilitate the exchange of test designs. Specific test designs are however not mandatory on the database CD-ROMs.

Five specific test designs for the Norwegian FDB are summarised in Table 1. The mnemonic TST is a number identifying each test. CCD is the set of SpeechDat(II) corpus codes included. The utterance content belonging to these corpus codes will be explained in Section 5. TSZ is the number of utterances in the test set. VSZ is the test vocabulary size and SCM is the scoring method. *Sentence match* scoring is intended for isolated word or phrase recognition, whereas *word match* scoring uses the NIST dynamic programming scoring algorithm. ACN is the number of words or sentences used in computing the scores. In all designs in Table 1, effectively empty utterances, as well as utterances with truncations, mispronunciations, unitelligible and out-of-vocabulary speech have been removed. A full test specification also includes a rejection count RCN. Since all potentially rejectable utterances are removed, this will be zero for all tests reported here.

Training corpus	# utterances	# triphones	# states before tying	# states after tying
VI-S	3980	6194	18576	1226
VI-OW	4777	5690	17064	514
VI-SOW	9215	8072	24210	1538
VD	6343	418	1248	397
Full	22588	8116	24342	2123

Table 2: Size and statistics for the five training corpora used

3. RECOGNISER DESIGN

The recogniser design is generally based on the tutorial example of the hidden Markov model toolkit (HTK 2.1 [8]). Instead of flat start initialisation, we use already existing models trained on the Norwegian EUROM.0, EUROM.1 and TABU.0 databases [9]. These models use 39-dimensional MFCC_E_D_A_Z features, which are not suitable for real-time recognition because of the utterance normalisation delay.

In the Norwegian SpeechDat database, there are 816 speakers with 45 utterances each available for training. Utterances with intermittent noise ([int]), truncated words (~), mispronunciations (*), unintelligible speech (**), filled pauses ([fil]), and phonetic letter pronunciations (/ /) were removed, leaving 22588 utterances in the *Full* trainset.

The original 46 Norwegian SAMPA phonemes occuring in the SpeechDat lexicon were reduced to 40 [5] and silence and tee models were added, as described in [8].

The training procedure is identical to the one described in [5], except that variance thresholds are used. These threshold values were set to 1% of the global trainset variance, as suggested in [8].

Training started from context independent, diagonal covariance single Gaussian HMMs with a standard three-state left-to-right topology. Pronunciation variants for each word in the trainset transcriptions were determined automatically by Viterbi alignment, and embedded Baum-Welch reestimations were used to update models.

From the monophone models, training proceeded by building word-internal context-dependent models for all triphones occuring in the training set. Decision tree state tying was then used to reduce the total number of HMM states and improve generalisation ability.

As a final training stage, the tied state triphone models were improved by mixture density modelling. Mixture models were generated by successive mixture splitting and reestimation. This resulted in a sequence of models with 2, 4, 6, 8, 12 and 16 mixture components, respectively.

4. TRAINING CORPUS SELECTION

In order to assess how different parts of the database contributes to recognition accuracy, models were trained from four different subsets, in addition to the Full trainset, as shown in Table 2.

Three of the subsets only contain phonetically rich material, without a number of word repetitions typically needed for whole-word model training. They can thus be considered vocabulary independent. *VI-S* consists of continuously read sentences from the corpus codes S1-9. The *VI-OW* set contains phonetically rich isolated words (W1-4) and names (O1-4,O7) only. *VI-SOW* is the union of the two sets above plus the sentence S0, which is an additional sentence, optionally included for speaker verification testing purposes.

The two remaining training sets contain repetitions of the vocabularies expected in test 01, 02, 03 and 04, and is therefore considered vocabulary dependent. *VD* consists of utterances with digits (I1, B1, C1, C4), application words (A1-6) and yes/no items (Q1-2), whereas the *Full* trainset contains all the VI-SOW and VD utterances, plus 14 additional utterances with spelled letters, natural numbers, time, date, money expressions etc.

Table 2 shows utterance counts, the number of word internal triphones encountered in the training set, as well as the total number of states in the triphone models before and after state tying. This gives an estimate of the relative complexity of models trained on the different training sets.

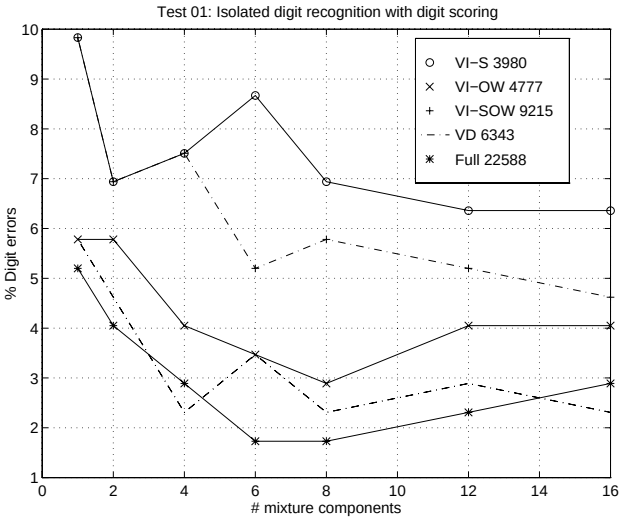

Figure 1: Isolated digit recognition results (digit scoring)

Training corpus	% Word errors	% Digit errors
VI-S	8.67	6.36
VI-OW	4.62	2.89
VI-SOW	6.94	4.62
VD	4.62	2.31
Full	4.62	1.73

Table 3: Best results on isolated digit recognition

5. TEST RESULTS

The training procedure described in Section 3 applied to one of the five training sets above in Table 2, leads to a set of acoustic models. In order to test these models on a given task, a test grammar and a pronunciation lexicon is needed. Since the SpeechDat lexicon covers all words occurring in the five tests. reported here, it could be used directly, with the phoneme mappings mentioned earlier.

Test 01 is isolated digit (0-9) recognition. The 12 vocabulary words used in the grammar includes synonyms for 1 ("en/ein") and 7 ("sju/syv"). Since most of the confu-

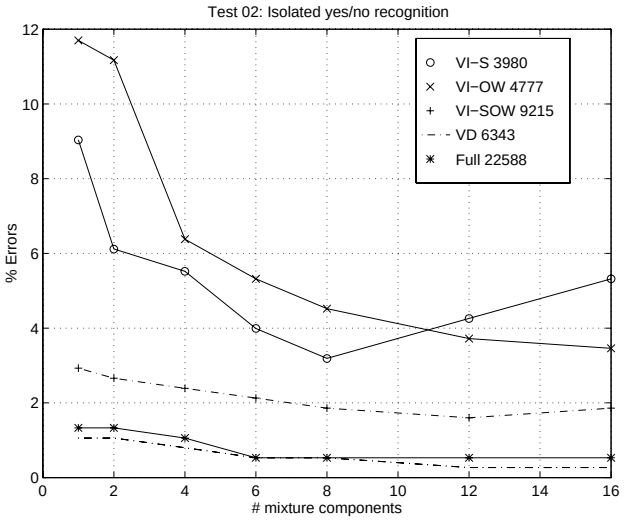

Figure 2: Yes/no recognition results

Training corpus	% Errors
VI-S	3.19
VI-OW	3.46
VI-SOW	1.60
VD	0.27
Full	0.53

Table 4: Best results on isolated yes/no recognition

sions are between "en" and "ein", we see from the results in Table 3 that digit error rates are significantly lower than word error rates for all training sets. From Figure 1 we see that the training sets including sentence material (VI-S and VI-SOW) perform worse than the one without (VI-OW). This may be explained by the fact that the digit word "en" is also a very common article in Norwegian, so that its models are likely to be corrupted by the many function word pronunciations in the sentences. Furthermore, we see that the Full trainset is slightly better than VD. This can be explained by the relatively large amount of digits present in natural-number items in the Full trainset. When comparing numbers, one should however remember that they are computed from 173 utterances only, which is not enough to give high confidence.

Test 02 is isolated yes/no recognition, with results shown in Table 4 and Figure 2. As opposed to test 01, the sentence-only trainset VI-S is now better than the word trainset VI-OW, and the combination VI-SOW is better than either of the two sets alone. The vocabulary-dependent models are also significantly better than all the VI models. This is the normal behaviour expected, considering the amount of phone model training data available in the sets.

Test 03 is isolated application word/phrase recognition. The vocabulary contains 12 one-syllable and 18 two-syllable command words and phrases, some highly confusable, like "slett/slutt". From the results in Figure 3 and Table 5 we have the same general observations as for test 02; more relevant training data improves performance. Here we also see that the extra material in the Full set (compared to VD) helps improve the high-complexity models.

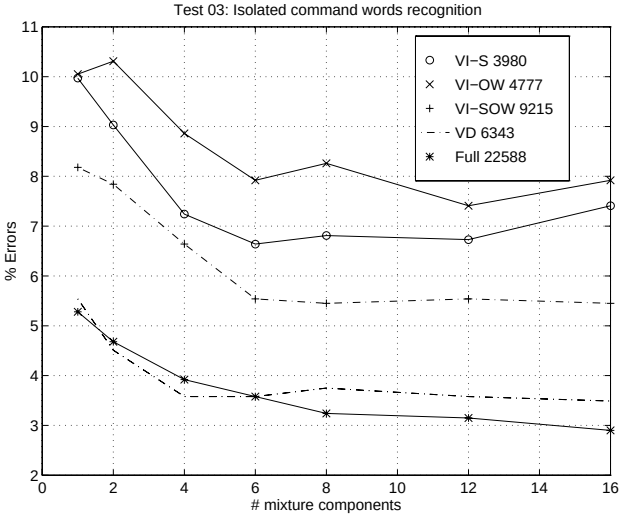

Figure 3: Application word recognition results

Training corpus	% Errors
VI-S	6.64
VI-OW	7.41
VI-SOW	5.45
VD	3.49
Full	2.90

Table 5: Best results on isolated application word recognition

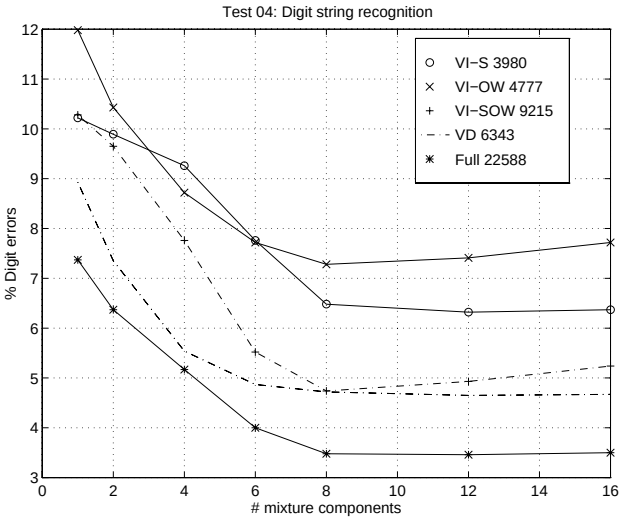

Figure 4: Digit string recognition results

Training corpus	Word scoring		Digit scoring	
	word error	string error	word error	string error
VI-S	7.17	34.38	6.32	31.60
VI-OW	8.09	36.63	7.28	34.20
VI-SOW	5.78	28.65	4.74	23.78
VD	5.46	25.69	4.65	22.22
Full	4.35	20.83	3.46	17.01

Table 6: Best results on digit string recognition

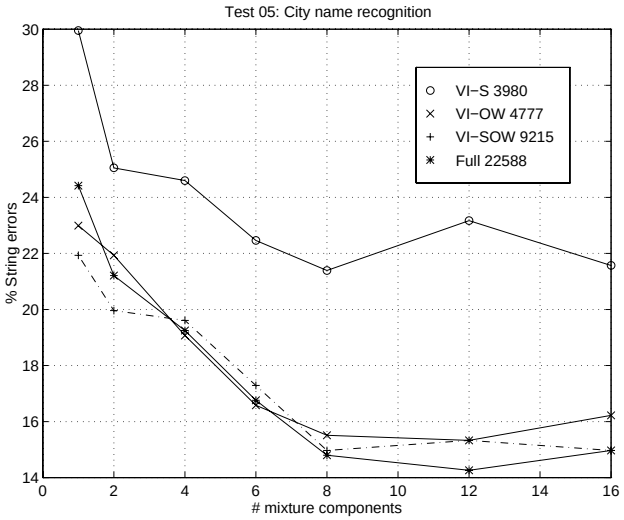

Figure 5: City name recognition results

Training corpus	% Word errors	% String errors
VI-S	21.76	21.39
VI-SOW	14.95	14.97
VI-OW	15.28	15.33
Full	13.95	14.26

Table 7: Best results for city name recognition

Test 04 is unknown length digit string recognition. The test material contains strings of length 6, 8 and 10, presented to the speaker in different formats. In the 10-digit string, speakers were instructed to include pauses between each digit, and this leads to more insertion errors than the connected digit reading mode. The recogniser used a phone-loop grammar with a logarithmic word insertion penalty of -100. The results in Figure 4 and Table 6 show that phonetically rich sentences now contribute positively, as opposed to the case in the isolated digit test.

Test 05 is city name recognition. The recogniser grammar contains 1143 city name phrases taken from all O2-4 items in the entire database (both training and test speakers). Results are given in Figure 5 and Table 7. The VD trainset is not included since the models trained by this set do not have enough phones to model all words in the test. We see that the significant difference observed is between the trainset containing phonetically rich sentences only and the others, which contain isolated word material. The overall best result on this task is 14% errors obtained with the full training set. A closer analysis of this result revealed that 10% of these errors were caused by heterographs in the transcriptions, i.e. different spellings for the same city name.

6. CONCLUSION

The results presented in this paper validates that it is feasible to build flexible vocabulary recognisers from a SpeechDat(II) 1000 speaker FDB, using standard HMM design methods. The results illustrate the importance of having phonetically rich isolated word material in the corpus, not only sentences, as in Polyphone and SpeechDat(M).

For application word recognition, the error rate with vocabulary independent modelling is about one and a half times that obtained with vocabulary dependent phone modelling techniques. The introduction of extra speech material by using the full database for training, in addition to the vocabulary dependent material, does not hurt performance significantly in any of the tests.

The future direction of research will be to extend the experiments to different databases, with different languages and telephone networks, while keeping a common recogniser design. The current results on digit string and city name recognition indicate that this design also has potential for improvement. Flat start initialisation and real-time feature analysis are other obvious candidates for further study.

7. REFERENCES

[1] H. Höge, H. S. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. 1771–1774, Apr. 1997.

[2] A. Constantinescu, O. Bornet, G. Caloz, and G. Chollet, "Validating different flexible vocabulary approaches on the Swiss French PolyPhone and PolyVar databases," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pp. 2293–2296, 1996.

[3] H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini, and L. Boves, "Localizing an automatic inquiry system for public transport information," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pp. 853–856, 1996.

[4] U. Bub, J. Köhler, and B. Imperl, "In-service adaptation of multilingual hidden-Markov-models," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. 1451–1454, Apr. 1997.

[5] F. T. Johansen, "The Norwegian part of SpeechDat: A European speech database for creation of voice driven teleservices," in *Proc. IEEE Nordic Signal Proc. Symp. (NORSIG)*, pp. 101–104, June 1998.

[6] F. T. Johansen, I. Amdal, and K. Kvale, "The Norwegian part of SpeechDat: A European speech database for creation of voice driven teleservices," in *Proc. Norw. Signal Proc. Symp. (NORSIG)*, pp. 40–43, May 1997.

[7] G. Chollet, F. T. Johansen, B. Lindberg, and F. Senia, "Test set definition and specification," Tech. Rep. LE2-4001 – SD1.3.4, SpeechDat deliverable, 1998.

[8] S. Young, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book (for HTK Version 2.1)*. Entropic Cambridge Research Laboratory, Mar. 1997.

[9] K. Kvale and I. Amdal, "Improved automatic recognition of Norwegian natural numbers by incorporating phonetic knowledge," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. 1763–1766, Apr. 1997.