The COST 249 SpeechDat Multilingual Reference Recogniser

**Finn Tore Johansen, Narada Warakagoda (1), Børge Lindberg (2),
Gunnar Lehtinen (3), Zdravko Kačič, Andrej Žgank (4), Kjell Elenius, Giampiero Salvi (5)**

(1) Telenor R&D, Kjeller, Norway,
(2) Center for PersonKommunikation (CPK), Aalborg, Denmark,
(3) Swiss Federal Institute of Technology (ETH), Zurich, Switzerland,
(4) University of Maribor, Slovenia,
(5) Kungliga Tekniska Högskolan (KTH), Stockholm, Sweden,

Abstract

The COST 249 SpeechDat reference recogniser is a fully automatic, language-independent training procedure for building a phonetic recogniser. It relies on the HTK toolkit and a SpeechDat(II) compatible database. The recogniser is designed to serve as a reference system in multilingual recognition research. This paper documents version 0.95 of the reference recogniser and presents results on small and medium vocabulary recognition for five languages.

1. Introduction

Within the EU supported COST 249 action "Continuous speech recognition over the telephone", collaborative efforts have been devoted to the promotion and dissemination of speech recognition research results in a multilingual environment. The aim is to cooperate across tasks and languages to improve the state-of-the-art, and to focus on what are considered to be major research problems within speech recognition: robustness and multilinguality.

The first step in this work has been to set up procedures for reference recogniser training and benchmark testing, with a minimal dependency on the language, task and application. The procedures rely on databases in a standardised format being commonly available for a number of languages. With such a reference recogniser, large development efforts to build a state-of-the-art speech recognition system for each new language can be avoided. It is also possible to indirectly compare error rates between different tasks and languages.

The SpeechDat(II) databases and standards (Höge et al., 1999) were chosen as the source of multilingual speech data and as a standard regarding speech files, label files, annotation conventions, lexica, et cetera, since a significant number of SpeechDat(II) compatible databases are now available world-wide.

Previous work on reference recogniser development includes the now concluded projects COST 232, CAVE and PICASSO as well as the COST 250 project on "Speaker Recognition in Telephony". Also in (Schultz and Waibel, 1998) efforts towards handling of multiple languages are being presented. However, with the present SpeechDat-based reference recogniser, it is possible to make comparable results on a to this date unseen number of languages.

The automated procedure applied in this project creates a set of acoustic phoneme models directly from the SpeechDat(II) CD-ROMs using the language-dependent knowledge embedded in the database. It is based on a bootstrapping procedure that works without pre-segmented data. The procedure is based on the HTK toolkit (Young et al., 1997) and accounts for differences between languages and imperfections in database design. A test suite representing different typical applications is also included, along with a public web site for exchange of software and results.

In the following sections the SpeechDat(II) database design is briefly presented, followed by a description of the reference recogniser design and the benchmark test procedures. Finally results are presented and discussed.

2. The SpeechDat(II) databases

Within the SpeechDat(II) project (Höge et al., 1999) a total of 28 databases have been collected covering eleven European languages as well as some major dialectal variants and minority languages. 20 databases have been recorded over the fixed telephone network (FDB), 5 databases over the mobile network (MDB), and 3 databases have been designed for speaker verification via telephone (SDB). The recordings of the FDB and MDB databases cover between 500 and 5000 calls by different speakers being recorded in a single session (except for two MDBs using multiple sessions). The duration of each recording session was 4-8 minutes.

The databases are intended to be used for developing a number of applications such as information services (e.g. timetable information), transaction services (e.g. home shopping, home banking) and other call processing services.

The three different types of SpeechDat(II) databases (FDB, MDB and SDB) share a core of roughly 40 utterance types as shown in Table 1.

SpeechDat(II) compatible databases have been or are being recorded in SpeechDat(E) (FDBs for five central and eastern European languages), SALA (SpeechDat across Latin America), SpeechDat-Car (SpeechDat databases recorded in a car environment) and for Australian English (Höge et al., 1999).

3. Recogniser design

The reference recogniser training procedure is an extension of the HTK tutorial example in (Young et al.,

number	type	corpus code
1	isolated digit items	I
5	digit/number strings	B,C
1+	natural number(s)	N
1	money amounts	M
2	yes/no questions	Q
3+	dates	D
2	times	T
3	application keywords/keyphrases	A
1	word spotting phrase	E
5	directory assistance names	O
3	spellings	L
4+	phonetically rich words	W
9	phonetically rich sentences	S
40+	In total	

Table 1: Core utterance types for SpeechDat(II) databases

Pre-emphasis	0.97
Frame shift	10 ms
Analysis window	Hamming
Window length	25 ms
Spectrum type	FFT magnitude
Filterbank type	Mel-scale
Filter shape	Triangular
Filterbank channels	26
Cepstral coefficients	12
Cepstral liftering	22
Energy feature	C_0
Deltas	13
Delta-deltas	13
Total features	39

Table 2: MFCC_0_D_A feature set

1997). Decision-tree state clustered, word-internal context-dependent phonetic HMMs are trained from orthographic (word-level) transcriptions using a pronunciation lexicon and a "flat start" boot-strapping procedure. The training procedure is implemented mainly as a set of Perl scripts running on Unix platforms.

The training procedure starts by importing a SpeechDat(II) database. The SpeechDat files needed are the A-law speech files, the SAM format label files, the pronunciation lexicon file and the test session list file specifying a list of the 200 or 500 official test sessions, depending on the size of the database.

The test sessions are of course excluded from training. In addition to this, 10% of the training sessions are reserved for development testing. All subcorpora of different utterance types are included since this was found useful in (Johansen, 1998).

Individual utterances are discarded if their annotated content indicates the necessity. In particular, all utterances containing intermittent noise (marked as [int]), truncated recordings ($\sim$), mispronunciations (*), unintelligible speech (**), filled pauses ([fil]), and phonetic letter pronunciations (/ /) are removed from the training set. The noise markers for stationary noise ([sta]), speaker noise ([spk]), and filled pauses ([fil]) are currently ignored.

The acoustic features are conventional 39-dimensional MFCCs, including the zero'th cepstral coefficient C_0 as energy, as well as first and second order deltas, as specified in (Young et al., 1997) and summarised in Table 2. These features are suitable for real-time operation, but are not optimised to be robust for telephone speech or mobile phones in particular.

The SpeechDat(II) lexica are used to provide phonemic transcriptions for supervised training. Optional prosodic information is removed from the lexicon. A language-dependent phonetic mapping can optionally be specified to avoid modelling very rare phonemes.

Each (mapped) phoneme is modelled as a three state left-to-right HMM, with no skip transitions. Diagonal covariance Gaussians are used. Tied silence and tee models (models having a non-zero entry to exit transition probability) are added as described in (Young et al., 1997), to take care of both background noise and silence.

Training starts from context-independent, single Gaussian monophones. All Gaussians are initially boot-strapped unsupervised, to the global mean and variance of the training set. These "flat-start" models are then re-estimated by the supervised embedded Baum-Welch procedure. To reduce the problem with unlabelled silence between words, only the phonetically balanced sentences (subcorpus S1-9) are used in these boot-strapping stages. A Viterbi realignment is then performed on the whole training set. This allows lexicon pronunciations other than the canonical ones to be chosen and also identifies potentially erroneous annotations.

The initial monophone models are successively split and re-estimated into 2, 4, 8, 16 and 32 Gaussian mixture components. The 32-mixture monophones are used to segment the training set in another forced alignment. The obtained phoneme segment boundaries are then used to create entirely new monophones in an isolated-word training style. This two pass bootstrapping procedure was added in version 0.94, and improved performance significantly compared to the single pass bootstrap used in previous versions.

From the freshly initialised single-mixture monophone models, training proceeds by building word-internal context-dependent models for all triphones occuring in the training set. Word boundaries are modelled with left- or right-context-dependent models (biphones). The monophone models are first cloned, then re-estimated with context-dependent supervision.

In order to reduce the total number of HMM states and improve generalisation ability, state tying is performed. A top-down decision tree clustering approach ensures that unseen words can be modelled without retraining the models, as required for flexible vocabulary recognition. The clustering algorithm optimises the likelihoods of the training data by successively splitting nodes in a binary tree structure according to yes/no questions regarding phonetic context.

Decision tree questions are defined by a set of broad class definitions. In the current version, broad class definitions from five languages (Danish, English, Norwegian, Slovenian and Swiss German) are included. Since many of the SAMPA symbols are common between languages,

Language (database)	Train spkrs	Total uttr	Train uttr	Lexicon pronuns	Mono-phns	Tri-phns	State clstr
Danish FDB1000	800	34400	23216	39604*	71*	13056	7.3 %
Danish FDB4000	3500	150500	101100	39604*	71*	19032	11.5 %
Norwegian FDB1000	816	36720	20335	14826	40*	7866	8.4 %
Slovenian FDB1000	800	34392	20548	6011	39*	6613	10.8 %
Swedish FDB1000	800	38400	24827	25946	46	10689	8.6 %
Swedish MDB1000	800	41600	34346	16050	46	11876	7.8 %
Swiss German FDB1000	800*	32580	17442	30525	51*	12374	7.1 %

Table 3: Training statistics. A star (*) next to a number indicates that information external to the official SpeechDat database was used. This is either a session list, a pronunciation lexicon, or a phoneme mapping.

and since the decision tree will automatically select the best questions for the data, it makes sense to use the union of broad class definitions for new languages.

As a final training stage, both the fresh monophones and the tied state triphone models are once again improved by Gaussian mixture splitting and reestimation up to 32 components.

4. Test design

The models trained by the procedure above can be used to provide benchmark results for a number of different applications, in different languages. However, in order to analyse the general behaviour of the models between languages and to improve the general design, it is useful to have a commonly defined test suite based only on the SpeechDat(II) database itself.

For this purpose, the official SpeechDat(II) test sessions (Chollet et al., 1998) are used, with different subcorpora representing typical test applications. Five common tests have so far been designed, for some of the sub-corpora in Table 1:

- I-test: Isolated digit recognition (SVIP)
- Q-test: Yes/no recognition (SVIP)
- A-test: Recognition of 30 isolated application words (SVIP)
- BC-test: Unknown length connected digit string recognition (SVWL)
- O-test: City name recognition (MVIP)

For all these tests, common test procedures are used, ensuring identical rules of test design across databases and languages. Currently, there are three such procedures, denoted SVIP (Small Vocabulary Isolated Phrase), SVWL (Small Vocabulary Word Loop) and MVIP (Medium Vocabulary Isolated Phrase). Utterances with OOV, mispronunciation, unintelligible speech or truncations are excluded in all procedures, since these are difficult to score without a particular application dialogue in mind. Noise markers are also ignored. The standard US NIST algorithm (Young et al., 1997) is used when string alignment is needed for scoring, currently only in the SVWL test procedure.

To completely specify the details of a test, the test procedure, test corpus codes and test vocabulary must be selected for every database. The vocabulary may contain semantic mappings, to specify that synonym confusions (e.g. zero/oh in English) should not be counted as errors. In the MVIP test procedure, the vocabulary can also be generated automatically, from all (training and test) utterances in the database. If this is done, no utterances will be considered OOV.

In order to have a completely open experiment setting, both the training and test procedures and detailed test results are available on the public web site (COST 249 SpeechDat SIG, 2000). The intention of this site is to enable all researchers with access to a SpeechDat(II) database and the HTK toolkit to repeat the experiments reported and contribute to an improved common recogniser design.

5. Current results

Six different labs within the COST 249 community have successfully completed the training procedure (version 0.93) for SpeechDat(II) compliant 1000-speaker databases (Johansen et al., 1999). Here we present results for version 0.95, obtained for five languages. First, we summarise some training statistics in Table 3. More details are available on the web (COST 249 SpeechDat SIG, 2000).

We can see that the number of utterances actually used during training is significantly lower than the total number of utterances by the training speakers. Most of this reduction is due to the content filtering, and especially the removal of utterances with intermittent noise. The high number of monophone models for Danish is due to the fact that a large number of diphthongs are treated as phonemes in the lexicon. The rightmost column in Table 3 is the reduction in effective number of states obtained by the state tying procedure.

The various stages of the training procedure results in a relatively large number of models. A typical evolution of test results is shown in Figure 1, where the first-pass monophones are denoted “mini_M_N” (M is the number of mixture components and N the number of training iterations), the second-pass monophones models are “mono_M_N”, whereas “tri_M_N” and “tied_M_N” are triphones and state-tied triphone models, respectively. We see that the second-pass monophones with a low number of Gaussian mixture components are significantly better than the correspond-

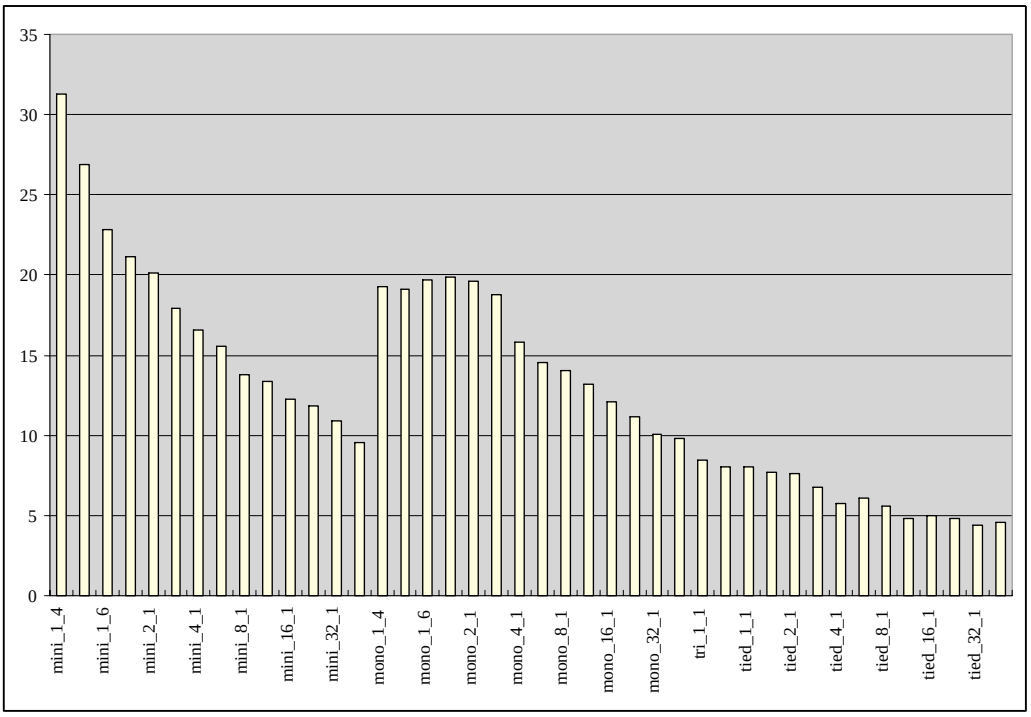

Figure 1: Typical performance evolution during training (Norwegian, A-test)

Language (database)	Test corpus				
	I	Q	A	BC	O
Danish FDB1000	1.04	1.14	2.36	2.30	15.82
Danish FDB4000	0.62	1.05	2.41	2.70	14.03
Norwegian FDB1000	2.31	0.53	4.43	5.87	17.31
Slovenian FDB1000	4.15	0.87	4.86	6.14	9.33
Swedish FDB1000	1.03	0.00	1.18	2.52	12.37
Swedish MDB1000	10.50	1.13	4.04	14.22	18.59
Swiss German FDB1000	0.51	0.27	1.06	3.10	6.29

Table 4: Word error rates (in %) with refrec 0.95

ing models trained in the initial pass, although the high-complexity monophones don't seem to suffer from poor initial segmentation.

A summary of test results for the different databases tried so far is given in Table 4.

It is worth noting that there is considerable variation between languages. Some of this can be related to the different noise levels in the telephone networks (as is illustrated by the difference between then Swedish FDB and MDB), while differences in vocabulary and phoneme sets obviously contribute significantly as well. In Table 5 and Table 6, average word lengths for the different tests are presented. For the O-test, the difference in test vocabulary seems to have a significant impact on the test results. The observed differences for the small vocabulary tests are however harder to explain by these numbers only.

When comparing the results in Table 4 to state-of-the-art recognisers, one should remember that whole-word modelling is not used. This will be a typical choice at least for digit recognisers. Cross-word context dependencies are not modelled either, and the degree of tying has not been

Language	I/BC	Q	A
Danish	2.64	2.00	4.57
Norwegian	2.85	2.00	4.60
Slovenian	3.85	2.00	6.52
Swedish	3.33	2.50	6.23
Swiss German	3.70	2.50	6.67

Table 5: Average number of phonemes in test vocabularies

Database	#Words	Phonemes/word
Danish FDB1000/FDB4000	495	6.52
Norwegian FDB1000	1182	7.34
Slovenian FDB1000	597	10.36
Swedish FDB1000	905	9.29
Swedish MDB1000	869	8.96
Swiss German FDB1000	684	12.64

Table 6: O-test vocabularies

optimised. Furthermore, a lot of training data containing noise has been removed. Thus, the silence modelling is probably not good enough. There are still no models for speaker noise or filled pauses either.

6. Conclusion

The reference training and test procedures presented have shown to provide a solution to the practical/logistic problems for the languages involved so far. The web page set up to coordinate experiments will hopefully contribute to significant progress in research on SpeechDat(II) databases, by disseminating benchmark procedures and results obtained for different databases.

Future work will include the improvement of language and database coverage. So far mainly the smaller SpeechDat(II) databases have been applied for training and testing (1000 speaker databases) with the exception of the Danish 4000 speaker database.

Also the number of standardised tests will be extended to cover some of the more challenging tasks represented within the SpeechDat(II) databases. This includes large vocabulary tasks such as name recognition (company and/or person names).

With the standardised description of speaker demographics it is also possible to develop common error analyses on existing tests and thus to correlate performance figures with demographic information.

Finally, the aim is to find a more general approach to broad class partitioning and phoneme mapping, as this is currently the only linguistic information needed by the training procedure that can not be found in the database. This would make the reference recogniser more universal and ease the inclusion of new languages in the covered set of languages.

7. Acknowledgements

The COST 249 action and all its participants are acknowledged for their support and encouragement in relation to the work presented.

8. References

Chollet, G., F.T. Johansen, B. Lindberg, and F. Senia, 1998. Test set definition and specification. Technical Report Deliverable SD 1.3.4, SpeechDat project LE2-4001.

COST 249 SpeechDat SIG, 2000. The Refrec homepage. http://www.telenor.no/fou/prosjekter/taletek/refrec.

Höge, H., C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders, and H.S. Tropf, 1999. SpeechDat multilingual speech databases for teleservices: Across the finish line. In *Proc. Europ. Conf. Speech Proc. and Techn. (EUROSPEECH)*.

Johansen, F.T., 1998. Phoneme-based recognition for the Norwegian SpeechDat(II) database. In *Proc. Int. Conf. Spoken Language Processing (ICSLP)*. Sydney.

Johansen, F.T., B. Lindberg, G. Lehtinen, Z. Kačič, B. Imperl, B. Milner, D. Chaplin, K. Elenius, G. Salvi, E. Sanders, and F. de Wet, 1999. The COST 249 SpeechDat multilingual reference recogniser. COST 249 MCM, Technical Annex, Budapest.

Schultz, T. and A. Waibel, 1998. Language independent and language adaptive large vocabulary speech recognition. In *Proc. Int. Conf. Spoken Language Processing (ICSLP)*. Sydney.

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