Free Acoustic and Language Models for Large Vocabulary Continuous Speech Recognition in Swedish

Niklas Vanhainen and Giampiero Salvi
KTH, School of Computer Science and Communication,
Department of Speech Music and Hearing, Stockholm, Sweden
niklasva@kth.se, giampi@kth.se

Abstract

This paper presents results for large vocabulary continuous speech recognition (LVCSR) in Swedish. We trained acoustic models on the public domain NST Swedish corpus and made them freely available to the community. The training procedure corresponds to the reference recogniser (RefRec) developed for the SpeechDat databases during the COST249 action. We describe the modifications we made to the procedure in order to train on the NST database, and the language models we created based on the N-gram data available at the Norwegian Language Council. Our tests include medium vocabulary isolated word recognition and LVCSR. Because no previous results are available for LVCSR in Swedish, we use as baseline the performance of the SpeechDat models on the same tasks. We also compare our best results to the ones obtained in similar conditions on resource rich languages such as American English. We tested the acoustic models with HTK and Julius and plan to make them available in CMU Sphinx format as well in the near future. We believe that the free availability of these resources will boost research in speech and language technology in Swedish, even in research groups that do not have resources to develop ASR systems.

Keywords: Automatic Speech Recognition, Acoustic Models, Language Models

1. Introduction

Developing speech technology knowingly requires a wide range of expertise including signal processing, machine learning, phonetics (ASR), computational linguistics (Language Modelling), artificial intelligence (Dialogue Systems). In small countries, and for languages with a small pool of speakers, this range of competences can seldom be gathered in the same research group, or even in the same university or research institute. It is, therefore, necessary that linguistic resources be shared in order to stimulate research in this field.

In recent years, many free ASR software packages have been made available to the community. Examples are CMU Sphinx[1] (Lamere et al., 2003), Julius[2], and Kaldi[3] (Povey et al., 2011). These provide the algorithmic implementations required to run a large vocabulary recogniser, and, in the case of Sphinx, even some procedures to train acoustic models from speech data. However, as all systems based on machine learning, these algorithms are useless if the problem and language specific resources in the form of acoustic and language models are not provided.

Traditionally, the speech community has been, to use an euphemism, parsimonious in sharing speech resources, mainly due to the large costs involved with collecting good quality corpora. With the exception of the efforts in the VoxForge project[4], speech databases are usually distributed against a fee, even for research purposes, and with rather restrictive licenses that prevent researchers from sharing the results of their work (so called "derived work").

Recently, however, due to bankruptcy of the company that collected them (Nordisk Språkteknologi, NST), reasonably large speech and text corpora were released to the public domain for three Scandinavian languages: Swedish, Norwegian and Danish. These, together with N-gram language models for the three languages, are available for download at the Norwegian Language Council website[5].

Using the NST Swedish speech database, we revised the training procedure originally defined in RefRec (Lindberg et al., 2000; Johansen et al., 2000) for the SpeechDat telephone speech databases, and we trained acoustic models in a number of conditions. We made, then, these models available for free download and use at our website[6]. Additionally, we created a free plugin for the popular speech analysis software WaveSurfer[7] (Sjölander and Beskow, 2000) that makes the use of these models accessible to non expert users and that is detailed in Salvi and Vanhainen (2014).

Here we describe the modifications made to RefRec and a number of experiments for isolated word and continuous large vocabulary speech recognition. The results are compared with those obtained for the same task with the SpeechDat models and with results obtained in similar conditions for American English. Although this paper does not report on advances in speech recognition methods *per se*, to our knowledge, this is the first time that large vocabulary speech recognition results are reported for Swedish and it is the first time that good quality acoustic models for Swedish LVCSR are made freely available to the community.

2. The NST Databases

The NST databases were developed originally by the Norwegian company Nordisk Språkteknologi (NST) which went bankrupt in 2003. A consortium of Norwegian universities and international companies acquired the data and made it available through the Norwegian Language Council.

[1] http://cmusphinx.sourceforge.net
[2] http://julius.sourceforge.jp
[3] http://kaldi.sourceforge.net/
[4] http://www.voxforge.org/
[5] http://www.sprakradet.no/
[6] http://www.speech.kth.se/asr
[7] http://wavesurfer.sourceforge.net/

set	# speakers	# recordings	length (hours)
training	965	307568	420.8
test	76	73046	103.4

Table 1: Information about the NST speech database

The resources that are available for Swedish are a number of speech databases for speech recognition and synthesis, a lexical database for morphological and pronunciation modelling, a text corpus for language modelling and a set of N-gram statistics estimated on it. All the documentation is available in Norwegian at the Norwegian Language Council website and originates from a report written after NST bankruptcy (Andersen, 2005). For simplicity, the resources that are relevant to this paper are described in more details in the following.

2.1. Speech Databases

The Swedish corpora were collected for a dictation task in office environment, based on phonetically rich sentences extracted from the Swedish text corpus also available from NST. In addition to the news material contained in the text corpus, isolated words, number sequences, and spellings of words were also included in the recording prompts. The following information is partly obtained from Andersen (2011a) (in Norwegian) and partly extracted from the database files when the documentation does not correspond to the actual data. There are three sets of speech data: i) a database for speech recognition and dictation, ii) a database specifically designed for dictation and iii) a database of speaker noises. Of these, only the first was used in this study. The speech recognition database consists of training and test sets with the characteristics described in Table 1.

Recordings are made with two channels, 16kHz sampling rate, 16 bits resolution and stored in NIST format. Each recording session is augmented by a text file containing meta-information. This includes a transcription of each utterance, information on the recording equipment and on the speaker. In the documentation it is mentioned that during validation of the material, additional information was added to the annotations. This includes a quality assessment of the recordings and annotations of a number of non-linguistic and noise events, including eight different types of noises. Unfortunately, these improved transcriptions are not available for download, and, as the time of writing, we could not get access to them.

Although the documentation does not provide aggregate statistics about the speakers, from the transcription files we could extract the following: There is a total of 1041 speakers in the database, balanced for gender. The training and test sets are described in Table 1. The speaker age is in the range 18-70 and they are grouped into 10 different dialectal areas: Stockholm area, South East Sweden, South West Sweden, Västergötland, West Sweden, Östergötland, Dalarna and surroundings, Göteborg area, Middle Sweden and Norrland.

2.2. Lexical Database

The lexical database for Swedish is described in Andersen (2011b) (in Norwegian). The lexicon consists of one single text file. Each line in the file corresponds to an entry in the lexicon and consists of 51 fields separated by semicolons. Describing the content of each field is outside the scope of this paper. There are 927,167 words in the lexicon with at least one pronunciation each. Of these 0.67% have two pronunciation variants, 0.02% three pronunciations and 0.01% four pronunciations. About 100,000 entries are taken from the list of most frequent words encountered in the text corpus described below. The rest are automatically generated inflections, of which, 249,901 were manually checked. Although the lexicon contains information about stress, syllable boundaries and sub-word boundaries of compound words, these markers were discarded in our experiments.

A difference between the NST and SpeechDat lexicons is that the open allophones of /ɛː/,/øː/,/ɛ/,/ø/ before an /r/ (/æː/,/œː/,/æ/,/œ/), do not exist in the NST lexicon, and were therefore not used to train the models described here. In total the lexicon uses 41 phonetic symbols in the pronunciations.

2.3. N-gram Data for Language Modelling

Knut Hofland at Uni Research AS created N-gram statistics up to the 6-grams for the text corpus available for download at the Norwegian Language Council, and made them freely available. The text corpus contains text from newspapers, novels and magazines, and the statistics comprise a total of 4,238,495 uni-grams, 50,478,732 bi-grams and 166,094,677 tri-grams.

3. The Recogniser Design

The training procedure for the speech recogniser is similar to the one defined by the reference recogniser (RefRec, Lindberg et al. (2000; Johansen et al. (2000)). This is a set of Perl scripts that rely on the tools provided by the HTK, Hidden Markov Model Toolkit (Young et al., 1997) and are designed to manage the formats in the SpeechDat telephone speech databases (Höge et al., 1999). The procedure generates context independent and word-internal context dependent phonetic hidden Markov models (HMMs). Each models is a three state left-to-right HMM.

Only word level (orthographic) transcriptions are used in the training and model initialisation is "flat start" assigning the global means and variances of the data to each state. A first set of context independent models is trained using embedded Baum-Welch estimation and the Gaussian components are successively split into 2, 4, 8 16 and 32 mixture components. These models are then used to realign the speech material to a phonetic transcription obtained using the lexicon and a new phoneme dependent initialisation is performed. Finally new context independent and dependent models are trained. The model parameters in the context dependent models are clustered at the state level using a decision tree and the corresponding parameters are tied.

3.1. Modifications to RefRec for the NST Database

We adapted the RefRec procedure to the NST database formats. A set of scripts were written to manage the special formats used in the speech database transcriptions and in the lexical database. A number of extra features were introduced: i) the possibility to down-sample the speech ma-

terial and, therefore, train both on 16 kHz and 8 kHz data; ii) more flexibility in the choice of features to extract from the speech data, see Section 4. for details. iii) the Gaussian splitting procedure is extended to include 64 Gaussian components per state (GMM).

We implemented two kinds of tests for the acoustic models. The first is an *isolated word recognition test* that was derived from the Medium Vocabulary Isolated Phrases (MVIP) test in RefRec. It uses HTK Viterbi decoder in the tool HVite. A subset of the NST test data is selected so that each recording contains an isolated word. Words with identical pronunciation from the lexical database are clustered (usually these correspond to different spellings of the same name or surname). The aim of this test is to evaluate the performance of the acoustic models alone, without interference of the language models.

The second evaluation procedure is a *large vocabulary word loop test*. This was designed from scratch because no such test existed in the RefRec package. The test uses Julius two-pass decoder in combination with the language models we created from the text statistics described in Section 2.. In the first pass, the decoder uses a bi-gram model and in the second pass, a tri-gram model is used. Although the results of this test are strongly influenced by the quality of the language models, its purpose is to evaluate if the acoustic models can be used in a real-world scenario.

3.2. Baseline

To our knowledge, this is the first time large vocabulary ASR results are reported for Swedish. In the attempt to define a meaningful baseline, in this paper, we have compared our results we those we can obtain on the same tasks with the models trained on the SpeechDat database. This is a reasonable comparison because SpeechDat, among the Swedish databases we have access to, is the closest corpus in terms of size. However, SpeechDat is different from NST in many respects. Firstly, it is recorded over the telephone line. Although in the comparison we have down-sampled the NST corpus to 8 kHz both for training and testing, there may still be a channel mismatch that can severely influence the results. Secondly, the linguistic content is different between the two databases, SpeechDat containing mostly command words and phrases, whereas NST containing phrases from newspaper text. This may affect the distribution of context dependent models and, therefore, the results.

If language match is not the main concern, a database that is similar to NST is the Wall Street Journal (WSJ) database for American English (Doug and Baker, 1992). In (Vertanen, 2006) a baseline recogniser for WSJ was proposed that closely resembles the models trained in this study. We will use the results from this reference as a cross-language comparison conscious that these do not represent state-of-the-art results fro WSJ. The reason is that we are not using state-of-the-art recognition methods in this study, and that results obtained on different corpora and languages should, in any case, be taken as only indicative.

Training database	Sampling Rate (kHz)	Features	# free states before clustering	# free states after clustering
SpeechDat	8	MFCC_0_D_A	48096	7602
NST	8	MFCC_0_D_A	52261	12199
NST	8	MFCC_0_D_A_Z	52273	13008
NST	8	PLP_0_D_A	52264	12320
NST	16	MFCC_0_D_A	52252	12490
NST	16	MFCC_0_D_A_Z	52273	13420
NST	16	PLP_0_D_A	52249	12577

Table 2: Number of free HMM states after clustering

4. Experiments

We trained a number of model sets in different conditions. The first parameter we varied was the sampling rate: we used the NST database original sampling rate of 16 kHz but we also down-sampled the material to 8 kHz to simulate recognition across the telephone line. The latter allowed us to compare the results with the SpeechDat models that were trained on telephone speech. We trained models based on Mel Frequency Cepstral coefficients (MFCCs) with zero coefficient and first and second order time derivatives (MFCC_0_D_A[8]), also we trained models based on MFCCs with Cepstral mean subtraction (MFCC_0_D_A_Z). Finally we trained models based on Perceptual Linear Prediction (PLP) coefficients with zeroth coefficient and time derivatives (PLP_0_D_A).

The available training data was split in a developmental test of 14666 and a training set of 292902 utterances. The training procedure removed a number of utterances that were considered outliers. This number varied depending on the training parameters (feature set and sampling rate). The final training set contained between 252752 and 252845 utterances. In the training set there are 17388 within word triphones, whereas in the lexicon the number of triphones is 26242. As a comparison, the number of triphones in SpeechDat training set and lexicon were 16000 and 16742, respectively. The state clustering procedure depends on the model fit to the data and therefore on the sampling rate and set of features. Table 2 summarises the number of free states before and after clustering for the different NST trainings and for the SpeechDat models. From the table it can be seen that the NST database affords estimating more free parameters than the SpeechDat database.

We ran the two tests described in Section 3.. In the *isolated word recognition test* the number of recordings in the NST test set containing isolated words is 3200 corresponding to a vocabulary of 1398 words. Of the latter, 1378 have distinct pronunciations in the lexicon.

In the *word loop test* we used a subset of the NST test set described below. We used the N-gram data described in Section 2. to create language models that contain only the words in our lexicon. The final language model consists of 6,777,643 bi-grams and 50,020,409 tri-grams.

As in the training phase, we select only the utterances where the words are in the lexicon, we also chose utterances where the words existed in the language models. Fur-

[8] In the remainder of this paper we will use HTK codes to specify feature kinds.

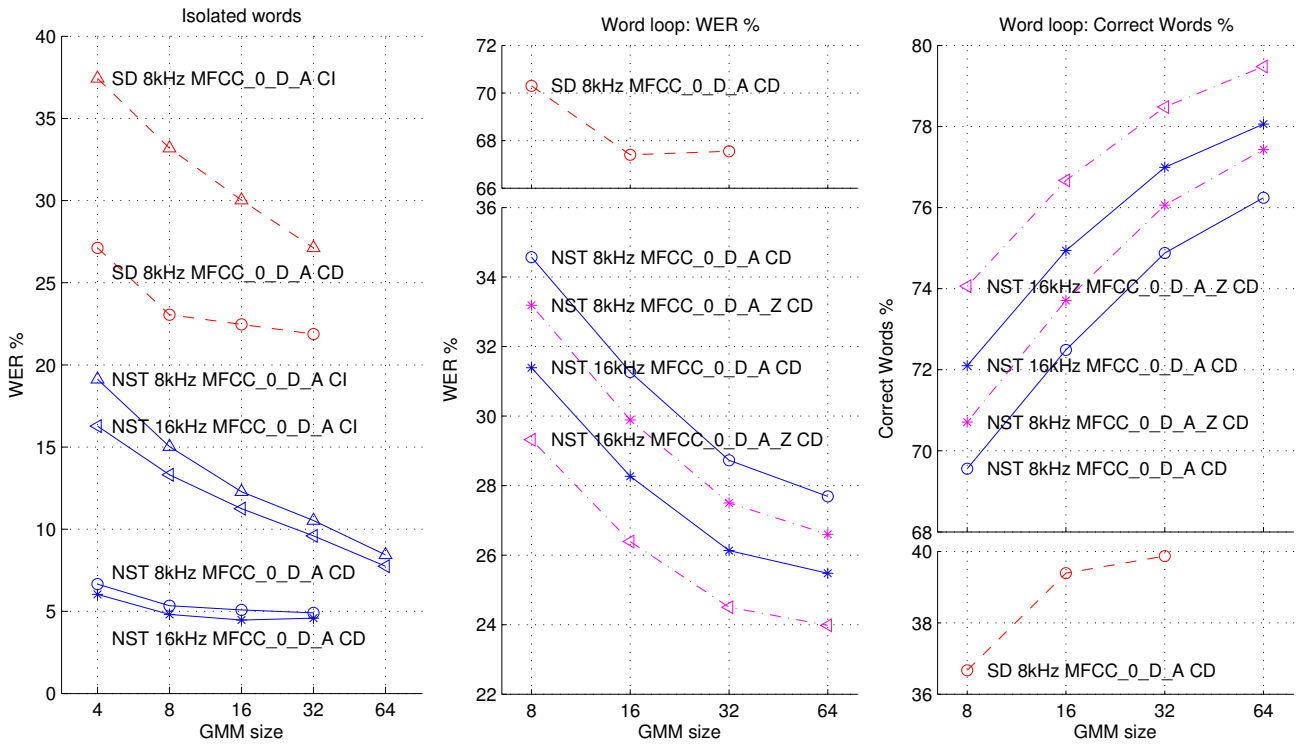

Figure 1: Summary of results for the models based on MFCC features. Left: word error rates for the isolated word recognition test. Middle: word error rates for the word loop recognition test. Right: percent correct word for the word loop test. SD = SpeechDat, CI = Context Independent, CD = Context Dependent acoustic models.

a) Isolated Word Recognition

Training database	Sampling Rate (kHz)	Features	Context	WER (%)
SpeechDat	8	MFCC_0_D_A	CI	27.12
SpeechDat	8	MFCC_0_D_A	CD	21.88
NST	8	MFCC_0_D_A	CI	8.44
NST	8	MFCC_0_D_A	CD	4.91
NST	8	MFCC_0_D_A_Z	CI	12.06
NST	8	MFCC_0_D_A_Z	CD	7.34
NST	8	PLP_0_D_A	CI	9.28
NST	8	PLP_0_D_A	CD	**4.88**
NST	16	MFCC_0_D_A	CI	7.75
NST	16	MFCC_0_D_A	CD	4.47
NST	16	MFCC_0_D_A_Z	CI	9.06
NST	16	MFCC_0_D_A_Z	CD	6.75
NST	16	PLP_0_D_A	CI	8.22
NST	16	PLP_0_D_A	CD	**4.44**

b) Word Loop Recognition (CD models)

Training database	Sampling Rate (kHz)	Features	Corr (%)	WER (%)
SpeechDat	8	MFCC_0_D_A	39.87	67.55
NST	8	MFCC_0_D_A	76.24	27.69
NST	8	MFCC_0_D_A_Z	**77.43**	**26.60**
NST	16	MFCC_0_D_A	78.06	25.48
NST	16	MFCC_0_D_A_Z	**79.48**	**23.98**

Table 3: Summary of Best Results

thermore, in order to make the test set more comparable to published LVCSR results for English, we excluded from the test set those utterances containing number sequences, spellings and isolated words that were added to the spoken database and that do not correspond to the text corpus the language models are trained on. The resulting test set consists of 31302 utterances. The recognition vocabulary contains 19763 words, of which 19337 have distinct pronunciation in the lexicon. To give an idea of the difficulty of the task, the perplexity obtained on the test set using the language models was calculated as 478.73 for the bi-gram model and 260.29 for the tri-gram model.

5. Results and Discussion

The results for the *isolated word recognition test* are shown in Figure 1 (left) and Table 3 (a) in terms of percent word error rates (WER %). In Figure 1 (left), results obtained with the SpeechDat models are compared to the ones obtained with the NST models using MFCC features. To make the comparison possible, both models trained on the original NST database (16 kHz) and the down-sampled version of the database (8 kHz) are shown. The NST models clearly outperform the SpeechDat models on this task. Models trained on the full spectrum (16 kHz) perform better than the 8 kHz counterparts, although this difference is more evident for Context Independent (CI) models than for Context Dependent (CD) models.

Table 3 (a) compares the performance obtained with MFCCs with and without Cepstral Mean Subtraction (CMS) and with Perceptual Linear Predictive coefficients (PLPs). For simplicity, only the best results are given for the different conditions. Here we can observe that CMS (MFCC_0_D_A_Z) causes a degradation in performance. This is probably due to the fact that, given the short length of the utterances used in this test, the normalisation

removes not only characteristics of the channel but also discriminative information about the words. Models trained on PLP coefficients (PLP_0_D_A) obtain the best results (in bold in the table), but the difference is marginal compared to MFCC_0_D_A.

Results for continuous speech recognition (*word loop test*) are shown in Figure 1 (middle and right) and in Table 3 (b) in terms of WER %, where insertions, deletions and substitutions are considered, and percentage of correct words. Again, the NST models clearly outperform the SpeechDat models (note that the y-axis is broken to better display results in different ranges). This time, however, the best results are obtained with MFCCs with Cepstral Mean Subtraction. The best performance in our tests is obtained with 64 Gaussian components per state and results in 23.98% WER and 79.48% correct words. Differently from the *isolated words test*, here the performance keeps on increasing when we add more Gaussian components, and is likely that improvements can be still achieved with this simple method. Note that, although the PLP features obtained the best results in the isolated word test (HTK decoder), the same tests could not be run for the word loop case because Julius only supports MFCC-derived features.

As a cross-language comparison, the baseline recogniser for the WSJ American English database described in (Vertanen, 2006) achieves word error rates between 20 and 25% on the San Jose Mercury sentences with models of comparable complexity as the ones in this study.

6. Conclusions

We have presented a collection of acoustic models that can be freely downloaded and used for large vocabulary speech recognition in Swedish. The best performance for continuous speech recognition is obtained, in this study, using context dependent models trained on Mel Frequency Cepstrum Coefficients with Cepstal Mean Subtraction. These models give a word error rate of about 24% and about 79% correct words. Although no previously published figures exist on a similar task for Swedish, these results are comparable with what reported for other languages in similar conditions.

7. Acknowledgements

This research has been partly supported by the Gothenburg Centre for Language Technology (CLT).

8. References

Andersen, G. (2005). Gjennomgang og evaluering av språkressurser fra NSTs konkursbo. Technical report, Norwegian Language Council.

Andersen, G. (2011a). Akustiske databaser for svensk. Technical report, Nasjonalbiblioteket.

Andersen, G. (2011b). Leksikalsk database for svensk. Technical report, Nasjonalbiblioteket.

Doug, P. B. and Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the DARPA SLS Workshop*, February.

Höge, H., Draxler, C., van den Heuvel, H., Johansen, F., Sanders, E., and Tropf, H. (1999). Speechdat multilingual speech databases for teleservices: Across the finish line. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*.

Johansen, F. T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G. (2000). The COST 249 SpeechDat multilingual reference recogniser. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., and Wolf, P. (2003). The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, pages 2–5. Citeseer.

Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recogniser based on SpeechDat(II). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *Proc. ASRU*, pages 1–4.

Salvi, G. and Vanhainen, N. (2014). The WaveSurfer Automatic Speech Recognition Plugin. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *LREC*. European Language Resources Association (ELRA).

Sjölander, K. and Beskow, J. (2000). Wavesurfer - an open source speech tool. In *Proc. of Interspeech*, pages 464–467.

Vertanen, K. (2006). Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Technical report, Cavendish Laboratory.

Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., (1997). *The HTK Book*. Entropic Cambridge University Laboratory, dec.