

SALTcymru_

Document 3

An overview of the Sphinx speech recognition system, and its possible enhancement for use in Wales in a pre-competitive research stage

**Prepared by the
Language Technologies Unit (Canolfan Bedwyr), Bangor University**

April 2008



Llywodraeth Cynulliad Cymru
Welsh Assembly Government

This document was prepared as part of the SALT Cymru project, funded by the Welsh Assembly Government under the Knowledge Exploitation Fund's Knowledge Exchange Programme, reference HE 06 KEP 1002

Introduction and relevance to the field

Speech recognition has been found to be a key area for the SALT Cymru survey. It was the second most popular area of SALT in terms of current development, and was considered very important or fairly important by 47% of developers. Additionally, 41% of developers stated that they might develop speech recognition in the future.

In searching for SALT that might aid Welsh SMEs, which are nevertheless able to be developed within a non-state aid environment, a useful model is that of free software development. Such software is able to be developed by academic institutions via grant or other funding, and collaboration in development is facilitated by access to the underlying source code. While the source code to any software can be considered an industrially valuable asset, certain open-source licenses (termed 'BSD-like') allow further commercial exploitation without exposing this commercially valuable source code. Hence, such licenses provide an ideal platform for developing in an environment that favours no one company or organisation, but nevertheless retains the ability for any interested parties to commercially develop any results that arise.

In developing speech recognition, it is not considered reasonable that a significant amount of work should be expended in coding basic algorithms that are well-known to the research community. It is recommended that use should be made of pre-existing packages, in which the algorithms are already coded, so that time can be spent developing applications within the existing frameworks.

The main packages for speech recognition, available at no cost, are **HTK** (developed originally by Cambridge University's Engineering Department) and **Sphinx** (a series of training and decoding decoders developed by Carnegie Mellon University). Of these, only Sphinx can be freely redistributed with end applications. It is free for commercial and non-commercial use, and has a BSD-like license. HTK would require the end-user to download and compile a version of HTK for their platforms.

In the majority of cases, it is not felt reasonable to ask an end-user to compile their own software to enable speech recognition. By this criterion, Sphinx should be the choice of recognition software.

Developing speech recognition for Wales

In developing speech recognition for applications in Wales, it is assumed that support is required for both the English and Welsh languages.

In English, speech recognition may be considered to be reasonably mature, as it has been the subject of development work for over 40 years. Of particular relevance to Sphinx is that acoustic and language models for English, trained on large amounts of data, are available for free download from an associated website¹. It is stated that this 'may just work' for individual needs, though it is implicit that additional data and development would be required for specialised applications.

¹ See <http://www.speech.cs.cmu.edu/sphinx/models/>

For Welsh, the situation is entirely different. Very little work has been done on speech recognition in the language, which means that development of any practical speech recognition system requires, essentially, its development from scratch.

Any speech recognition development involves the following stages:

1. Definition of a recognition task.
2. Selection and preparation of training and test data sets.
3. Training an acoustic model on the training data.
4. For all but the simplest recognition tasks, training a language model.
5. Running the decoder on the test data, and deriving a recognition score for the task.
6. Optionally, modifying the acoustic and language models to improve the recognition result on the given test data.

The description above details an off-line speech recognition system, i.e. one which operates on files of data rather than recognising live speech. If live speech recognition is required, the following must also be addressed:

7. Running the recognition stage on live speech input.
8. Optionally, adapting the acoustic and language models given the speech input data.

Definition of a recognition task

The recognition task comprises the range of speech the recogniser is expected to be able to deal with, in terms of the speaking environment(s), the range of speakers and the complexity of language. Recognition results are likely to be higher for a smaller range of speakers and a more restricted vocabulary within the application.

The recognition task must be carefully defined before further development takes place.

Selection and preparation of training and test data sets

In any speech recognition development, there is a trade-off between the amount of data used in training a recogniser and the quality of the eventual recognition result. A simple recogniser that distinguishes between a dozen or so words may only require a few minutes of training speech. At the other extreme of complexity, in most broadcast news recognition tasks over 100 hours of training speech are typically required to achieve about 30% word error rate in recognition.

In order to train acoustic models, the speech data must be segmented at the phonetic level. An automatic process of forced alignment will be used for this, involving individual utterances being aligned with a phonestring derived from their word-level transcriptions. Therefore, a pre-requisite for the forced alignment process is a phonetic transcription of each word in the training data. This will be derived from existing lexica developed for speech synthesis.

Forced alignment can begin once a transcription at the phoneme level is available for each word in the utterances. These transcriptions can be derived using freely available

phonetic dictionaries (lexicons) for Welsh, supplemented by the use of freely available letter-to-sound rules for those words not present in the lexicons.

Forced alignment works by taking an initial estimate of the segmentation of the utterances given their phonetic transcription. The initial estimate assumes that all phones in the utterances are of identical duration. Phone models are trained on this initial uniform segmentation, and then used to derive a new set of time-aligned transcriptions, which are found to be a slightly better match to the actual alignment than are the initial uniform segmentations. These new transcriptions are used to derive a further set of transcriptions, which are again found to be a closer match to the actual alignment. This process of convergence is repeated over a number of cycles to produce an accurate, automatically derived, time-aligned transcription of each utterance. Manual checking of a proportion of the results is essential, in order to determine

Training an acoustic model on the training data

Typically in speech recognition, acoustic modelling takes place at the phoneme level (i.e. at the level of the individual sounds of words). Individual Hidden Markov Models (HMMs) are used to model each phone in the language to be recognised. Standard techniques exist to train these models, and these techniques are implemented in Sphinx.

The only decision to be made at this stage of development involves the definition of a set of acoustic models for Welsh, i.e. the definition of the individual speech sounds which are to be modelled. Earlier work by the Language Technologies Unit at Bangor University has defined a set for the North Welsh accent². This however includes a large number of sounds. Depending on the amount of training data available to model each sound, it may be found advisable to use a simpler set of sounds, such as those used in the South Welsh accent. Any sounds present in the North Welsh accent would then be mapped to their equivalent in the South Welsh accent.

Training a language model

In speech recognition, a language model is used to reduce the number of possible candidates for the output text. It typically determines the probability of words in given contexts – usually, of a given word being followed by another given word or by a longer sequence of words.

A language model is trained from a large corpus of text suitable for the task in question. In the case of Welsh, two possibilities currently exist for this:

- CEG (approx. 1 million words)
- Crúbadán (approx. 95 million words)

While it would appear that the larger size of the Crúbadán corpus would result in a language model of higher quality, this is not the only criterion. The CEG corpus has been derived from primarily literary and newspaper texts. They have been quality-checked to a standard believed acceptable for most uses. Despite Crúbadán's larger size, its use may not result in a better recognition result. It has been collected from a broad selection

² See <http://bedwyr-redhat.bangor.ac.uk/svn/repos/WISPR/Documentation/Technical/phoneset-wispr-welsh.pdf>

of web text. As such it will probably contain many errors and text from other languages, notably English.

In the context of a language model, the number of errors in Crúbadán is less important than the type of those errors. Non-systematic errors in Crúbadán (e.g. specific words being largely spelt correctly, but occasionally suffering from typos) should not significantly affect the quality of the language model. Any n-gram based model would give a low probability to these errors, so they should not significantly influence the system's output. On the other hand, systematic errors in the corpus (e.g. a specific word being followed by another specific word, but incorrectly mutated) could affect the language model. If words occur regularly in specific incorrect contexts, the probability of those n-grams in the language model will consequently be higher, and the output of the system could be adversely affected.

In purely practical terms, the availability of CEG makes it more suitable for the rapid development of a language model in the first instance. However, it is strongly recommended that Crúbadán be investigated in future developments. It is suggested that after using a CEG language model within the speech recognition development, a Crúbadán language model should be substituted, and the results investigated.

Running the decoder on the test data, and deriving a recognition score for the task

At this stage in development, the acoustic and language models have been fully trained. A decoder can thus be selected, which will be given the trained models and input speech, and which will attempt to transcribe the text of the input speech.

Several versions of decoders have been developed as part of the Sphinx projects. They vary in methodology, in the programming language used to develop them, and in their levels of maturity. The most recently developed decoder is Sphinx-4, which is written in Java. However, it is aimed at off-line processing rather than live applications. Sphinx-3 is an older decoder, written in C and originally aimed at off-line processing, but which has recently been further developed for live recognition. It is regarded as the most accurate decoder developed as part of the Sphinx project.

The output of the decoding stage will be a set of transcriptions. To achieve a recognition score from these, a separate program is run. Sphinx includes a program that derives recognition scores which comply with the NIST standard, allowing results to be compared with speech recognition developments in other languages.

The decision of what constitutes an 'acceptable' recognition score is open to question. The state of the art in recognition of broadcast news bulletins is a word error rate of about 25% (a word recognition score of 75%). Such systems have normally been trained with 100 hours or more of speech data. However, any initial system for Welsh will have been trained on a much smaller amount of training data, and this should be borne in mind when comparing results.

Modifying the acoustic and language models to improve recognition of the test data

Modifications to acoustic modelling

A significant part of speech recognition development has involved adjusting various parameters within the recognition models and investigating their effect on the recognition result. The parameters to be used have by now been well-researched, and most speech recognition systems reported in research literature now use HMMs with three states per phone and MFCCs of about 13 orders with their first and second differential. However, some slight improvements in recognition results may be achieved by altering the topology of the HMMs – in other words, which transitions are allowed between their three states.

Varying the complexity of the acoustic models may also be necessary. A more complex acoustic model (technically, a greater number of Gaussian components per state of the HMM) allows a greater amount of training data to be modelled with increased accuracy. However, if there is not enough training data, then lessening the complexity of the acoustic models usually improves the recognition score.

Provided sufficient training data is available, the modification of the phone models to triphone models may improve recognition accuracy. Triphone models take into account the preceding and following contexts of the individual speech sound (the phone) being modelled. This results in an increased number of models and hence an increase in the amount of data required to train them. However, techniques exist to 'tie' the states of triphones with similar contexts, allowing the training data to be shared between two or more models. This reduces the total amount of speech data required to adequately train the models.

Modifications to language modelling

Analysis of the word patterns found within the text output may reveal errors and inconsistencies in the language model used. In particular, one unexplored area for Welsh is that of mutations and inflections, and how the text output would reflect those. It may be necessary to include mutations as alternative pronunciations of the same word form within the lexicon.

Other modifications

Additionally, the decoding stage may be made more accurate by increasing the word insertion penalty. In speech recognition, it is sometimes found that while all the words in the speech input are present in the text output, additional words have been introduced in error. These are termed 'insertion errors'. Increasing the word insertion penalty biases the decoder against including additional words in the output, and can help reduce the number of insertion errors.

General comments

It should be noted that to obtain new recognition scores requires that the whole process of training the language or acoustic models be repeated, the decoder re-run and a new recognition score derived. However, it is anticipated that by this point the process of doing so will have been automated through the writing of scripts, so it is expected that, beyond the actual modification of the models, the re-running of the training processes will be largely automatic processes.

Running the recognition stage on live speech input

The Sphinx 3 decoder can be set to run on live speech with no further integration or programming work needed. If any work is required in this section, it is anticipated that it will be in reformatting the output of the decoder, and piping or redirecting its output into additional programs.

It is anticipated that the main challenge of this stage of development (and indeed one of the main challenges in this development in general) will be in integrating the recogniser with a practical system. In particular, unlike the Festival speech synthesis system described in Appendix I1, Sphinx is not well-integrated with Microsoft Windows. Significant development work would have to be undertaken to ensure this integration, and it should be emphasised that this integration is not a trivial task.

Adapting the acoustic and language models given the speech input data

Standard techniques exist for acoustic model adaptation in speech recognition. The most common ones of these are MAP and MLLR. Sphinx supports both MAP and MLLR re-estimation. In particular, it allows MLLR adaptation to be performed on-line, while recognition is still taking place.

Techniques also exist for the adaptation of language models. This adaptation is beyond the scope of any initial development for Welsh, as it relies on the language model being updated off-line, and replacing the existing language model in the recognition stage once it has been updated.

General comments

It is evident from the above that the challenge of speech recognition for Welsh is a tractable and achievable one. However, it should also be implicit that previous expertise in the field is important in developing this complex area of SALT. This points towards the preservation of the existing skill base in Wales, and nurturing of future developers, as being essential in ensuring that this key technology is developed, and continues to be developed in future.