

# a Spoken Document Retrieval Application in the Oral History Domain

*Marijn Huijbregts, Roeland Ordelman and Franciska de Jong*

Department of Electrical Engineering, Mathematics and Computer Science,  
University of Twente, the Netherlands

{m.a.h.huijbregts, ordelman, fdejong}@ewi.utwente.nl

## Abstract

The application of automatic speech recognition in the broadcast news domain is well studied. Recognition performance is generally high and accordingly, spoken document retrieval can successfully be applied in this domain, as demonstrated by a number of commercial systems. In other domains, a similar recognition performance is hard to obtain, or even far out of reach, for example due to lack of suitable training material. This is a serious impediment for the successful application of spoken document retrieval techniques for other data than news. This paper outlines our first steps towards a retrieval system that can automatically be adapted to new domains. We discuss our experience with a recently implemented spoken document retrieval application attached to a web-portal that aims at the disclosure of a multimedia data collection in the oral history domain. The paper illustrates that simply deploying an off-the-shelf broadcast news system in this task domain will produce error rates that are too high to be useful for retrieval tasks. By applying adaptation techniques on the acoustic level and language model level, system performance can be improved considerably, but additional research on unsupervised adaptation and search interfaces is required to create an adequate search environment based on speech transcripts.

## 1. Introduction

The number of digital spoken-word collections is growing rapidly. Due to the ever declining costs of recording audio and video, and improved preservation technology huge data sets are created. But where the growth of storage capacity is in accordance with widely acknowledged predictions, the possibilities to index and access the archives created is lagging behind [1].

For content with a high economic value, such as news broadcasts, there is a wide range of techniques available for the generation of meta-data, including automatic speech transcription (ASR). Especially when the audience is large, as is the case for television broadcasts in English, the profitability of annotation stimulates investments in optimisation of tools for automatic metadata generation.

Less resources are spent on audio collections outside the news domain. Several interesting application domains exist, both within and outside the broadcast domain: documentaries, interviews, historical archives, recordings of lectures and recordings of meetings (corporate, scientific, parliamentary, etc.). Tools to browse in such collections and to search for fragments could support the information need of various different types of users, including archivists, information analysts, researchers, producers of new content, general public, etc.

A particular interest exists for the content of historical archives, which in combination with retrospective digitisation, represent a type of content which is rich in term of cultural

value, but which has a less obvious economical value. The observation that speech recognition can contribute to the disclosure of spoken word archives has been made many times [2], and several initiatives have been undertaken to develop robust spoken document retrieval (SDR) technology for audio collections in the cultural heritage domain ([3], [4]). Unfortunately, the high expenses required to process historical content in combination with the expected limited financial return on investment have prohibited real successes.

When resources to develop a speech recognition system for a new domain are not sufficiently available to enable the training of purely in-domain models, the most obvious step is to adapt an existing system. For example, given an oral history task domain with only few resources for training, these resources are deployed to tune a broadcast news system to the oral history domain. As generating additional training data manually is expensive, ideally, the tuning work using available data is supported by unsupervised adaptation techniques. These techniques should aim at the gradual improvement of acoustic model robustness using unseen data and the automatic selection of appropriate vocabularies and language models.

As a first step in our research trajectory towards such an adaptive system, an SDR application was attached to the web-portal that was officially launched in April 2005, dedicated to the Dutch novelist Willem Frederik Hermans (1921-1994) [5]. The data collection incorporates a number of different media types including text, images, audio and video and is very suitable for cross-media browsing [6]. For example, the site contains abstracts, biografies, information on novels, pictures, photographs, both written out and spoken interviews, lectures and 'spoken books'. The SDR system we implemented provides a facility for the indexing and searching of the audio and video content. Adding cross-media browsing facilities to the portal is currently being investigated.

Although the performance of a broadcast news (BN) system for the type of audio data encountered in the collection of the Willem Frederik Hermans (WFH) portal was expected to be poor, we used the resources and tools, collected and developed in earlier projects [7] for a BN system as a starting point. This system showed adequate performance in a Dutch spoken document retrieval task in the news domain. As similar systems are available in many labs, the conversion of the BN system and tuning to a collection from the oral history domain might be a case of a more general interest for research groups that want to pursue applications for their ASR tools for similar purposes. Although we are aiming for a system that can adapt to new domains unsupervised, we first investigated supervised adaptation methods.

First, in section 2 we outline the data collection, the available tools, and the methods used to optimise performance for the task domain. In section 3, we describe the different results

on the task specific evaluation data. In section 4 the results will be commented, and some consequences for future research will be mentioned.

## 2. Collection, systems and methods

### 2.1. WFH-collection

The collection of audio recordings to be disclosed consists of some 10 to 15 hours of lectures and interviews featuring Willem Frederik Hermans (WFH). More data will become available at a later stage. Although WFH is not the only speaker present in the material, his voice dominates the larger part of the collection. The lectures were recorded at different locations with different reverberation characteristics. The lectures which were studied in more detail have applause, laughter, coughing and questions from the audience that –even for a human listener– sometimes are hard to recognise. Parts of the interviews are quite informal and recorded in a home environment on celluloid tape.

### 2.2. Development and training data

One of the lectures and a television documentary with a number of interviews were manually annotated at word level, encompassing 130 minutes of speech, with WFH speaking approximately 85% of the time. This subset of the audio recordings was divided in a training set, a test set and an evaluation set. The training set (78 minutes) was used for training the acoustic models. The test set was used to evaluate both the acoustic models and the language models during development. The evaluation set was used for the final evaluation of the system.

For training BN language models we have collected a large Dutch news related text corpus of in total some 400M words [7]. Two other text collections were available for domain adaptation. A number of written interviews with WFH and one of his short novels made up the first collection (further referred to as WFH-text) containing one and a half million words. Word-level transcripts of general conversational speech from the Spoken Dutch Corpus [8] formed the other text collection. This collection consists of 1.65M words. Both text collections were used for adaptation of the language model.

### 2.3. Broadcast news system

Two broadcast news ASR systems were available as a starting point. The first system (UT-BN2002) is based on hybrid RNN/HMM acoustic models, a 65K vocabulary and a statistical trigram language model, created using a news corpus. The acoustic models are created out of approximately 20 hours of broadcast news speech.

The other system (UT-BN2005), is based on a recogniser which is developed at the University of Colorado (CSLR) and is freely available for research purposes. It uses decision-tree state-clustered Hidden Markov Models. [9].

Twenty-two hours of broadcast news recordings from the Spoken Dutch Corpus ([8]) were used to port the gender independent acoustic models from the English system to Dutch and to train new broadcast news acoustic models. The ARPA language model created for UT-BN2002 was reused in the UT-BN2005 system. The main advantage of UT-BN2005 over the UT-BN2002 system is that it allows to apply adaptation methods to the acoustic models, which is highly relevant given the mismatch between the target collection and the BN training material.

### 2.4. WFH system

The WFH transcription system is based on components of both BN systems. The BN language model and acoustic models were adapted to the target domain as will be described in this section.

#### 2.4.1. Language model adaptation

Two domain specific trigram language models were trained. These models both use a 30K vocabulary containing domain specific words.

The most occurring words from the WFH-text described above were complemented with the most occurring words from the newspaper corpus. Vocabularies of different sizes were created and the out-of-vocabulary (OOV) rates were computed on a preliminary test set. The OOV of a 60K vocabulary (50K from newspaper data and 10K from WFH-text) was only marginally better than the 30K vocabulary (20K newspaper words with 10K from WFH-text). As a smaller vocabulary will result in less acoustic confusability, the 30K vocabulary was chosen to be used for the WFH-specific language models.

From each of the two domain specific text collections described in section 2.2, a language model was created. The language model created from the WFH-text contains 178K trigrams and 335K bigrams. The conversational speech text collection contains 153K trigrams and 306K bigrams. A third model was created using the newspaper corpus and the 30K vocabulary. From these three models, a mixture language model was created. Mixture weights were computed using the transcripts of the acoustic training set.

#### 2.4.2. Acoustic model adaptation

In the annotated speech material, WFH is speaking 110 out of the 130 minutes. It is therefore reasonable to expect that the recognition rate will improve when a WFH-dependent acoustic model is used instead of the broadcast news acoustic model. Two new acoustic models were trained. Both models were evaluated using the mixture language model described in the previous section.

The first model was trained solely on the part of the training set in which WFH is speaking, in total 78 minutes of speech data. The second model was created by adapting the broadcast news acoustic model (UT-BN2005) to the training data using the Structured Maximum a Posterior Linear Regression (SMAPLR) adaptation algorithm [10]. In [9] and [10] it is shown that SMAPLR performs very good, even when little adaptation data is available.

#### 2.4.3. Dictionary

A pronunciation dictionary was created from the 30K vocabulary using a large background pronunciation lexicon provided by Van Dale Lexicography and a grapheme-to-phoneme (G2P) converter trained on the same pronunciation lexicon [7].

## 3. Experimental results

### 3.1. BN system performance

On the BN test set (4 hours from the Spoken Dutch Corpus), the UT-BN2005 system outperforms the BN-2002 system with word error rates (WERs) of 30% and 35% respectively. On the WFH test set both BN systems have a comparable performance of above 80% WER. The word error rate of the UT-BN2005 system is 80.4%. On the parts in which only WFH is speaking a

81.6% is scored, on other speakers we obtain a WER of 67.2%.

### 3.2. WFH system performance

To evaluate the performance of the WFH system, three evaluation runs were performed. First, the performance of the WFH language models will be reported. Next, the performance of the acoustic models will be discussed and finally, the performance of the combination of the best LM and AM will be reported.

#### 3.2.1. Language model

Table 1 shows the perplexities of the four created language models (see session 2.4.1), along with the word error rates on the test set of those models obtained using the BN2002 acoustic model. All language models use the same 30K vocabulary.

All language models perform better in terms of WER than the original news model. Merging all models into a single mixture model gave the best results on the WFH test set. Note that mixtures of two of the three language models did not improve results.

Name	PP	%WER
Newspaper	245	77.2
Conversational speech	274	78.8
Domain	235	77.1
Mixture	195	75.4

Table 1: *The perplexity (PP) and the word error rate (WER) of the four language models. All language models use the same 30K vocabulary. The WERs were calculated the 2002 broadcast news acoustic model.*

#### 3.2.2. Acoustic model

The word error rates of the broadcast news acoustic model and the adapted acoustic models are shown in Table 2. In order to make a fair comparison with the broadcast news system, the 65K broadcast news language model was used during these recognition runs. Table 2 shows three word error rates for each model. The first WER is of the part of the audio in which WFH is speaking, the second one is based on speech from other people and the third is the overall word error rate.

Both adapted models perform better than the broadcast news model. Although the broadcast news model performs best on the small subset with various speakers (15% of the total amount of speech), the adapted models show improved WERs on the part of the data in which WFH is speaking. The SMAPLR adapted model (66.9% WER) outperforms the speaker dependent model (76.6% WER). The 78 minutes of speech used for training the speaker dependent model does not contain enough data for building a robust acoustic model but using it for adaptation of the BN models does improve performance substantially.

#### 3.2.3. Overall results

The 30K mixture language model and acoustic models described in the previous sections were combined and tested on the evaluation set. Table 3 shows the word error rates of these combined systems.

The combination of the 30K mixture language model and the SMAPLR adapted acoustic model results in the best system performance: 66.9% WER.

AM	%WER	%WER	%WER
	WFH	other	total
UT-BN2005	81.6	67.2	80.4
WFH	76.0	83.2	76.6
BN/WFH	66.7	77.1	67.5

Table 2: *WERs of three acoustic models: the 2005 broadcast model, the WFH model and the SMAPLR adapted model. The second column shows the WER of the part in which WFH is speaking, the third the WER on the other speech parts of the evaluation set. The last column shows the total WER.*

Having a diarisation system available that could produce a series of time marks with associated speaker labels (“who spoke when”), the SMAPLR adapted system could be deployed for decoding the speech of WFH and the broadcast news system for recognition of other speech. Given a diarisation system that would produce our manual annotated segmentation, the word error rate would improve to 66.6%.

AM	%WER	%WER	%WER
	WFH	other	total
WFH	73.8	83.6	74.6
BN/WFH	66.4	72.5	66.9

Table 3: *The word error rates (WER) of the two adapted acoustic models combined with the 30K mixture language model. The first row contains the AM trained on WFH solely. The second row contains the SMAPLR adapted acoustic model.*

By creating a mixture LM and a speaker dependent AM the word error rate was reduced with 13.5% (16.8% relative). To determine possible further improvements, a brief error analysis was conducted as described in the next section.

### 3.3. Error analysis

To investigate to what extent different audio conditions influence the word error rate, speech segments were classified into five classes: clean speech ( $F_0$ ), speech with audible echo ( $F_1$ ), speech containing background music ( $F_2$ ), speech with background noise ( $F_3$ ) and overlapping speech (speech interrupted by other speakers,  $F_4$ ).

Table 3 shows the word error rates in each of the conditions. One third of the segments in the WFH evaluation set contains echo, music, noise or overlapping speech. Although all speech in the music class is clearly understandable for human listeners music increases WER substantially, in the WFH task by more than 10%, which is comparable with the statistics reported in [11].

Class	%WER	%WER	%WER
	WFH	other	total
$F_0$	63.9	61.8	63.8
$F_1$	75.4	100	76.5
$F_2$	76.1	86.1	78.6
$F_3$	82.4	83.3	82.4
$F_4$	100	100	100

Table 4: *The word error rates of the five manually classified parts of the WFH evaluation set.*

A stop-word list was used to filter out function words, hesitations and other words that are not helpful during search from

the speech recognition transcripts of the development set. Only 25% of all words remain after removing stop words (60% of all unique words). The out-of-vocabulary rate of the filtered text is higher than the original OOV (9.1%). It was hypothesised that these OOV words are typically domain specific words that might be learned from additional domain specific text material. Going manually through the OOV list though, revealed that only a small number of words and names can be regarded as domain specific words. The majority of the OOVs appeared to be either misspelled words, highly infrequent words or words that can not directly be related to the task domain.

#### 4. Summary and discussion

We reported the tuning of a Dutch large vocabulary speech recognition for a recognition and retrieval task in the oral history domain with a collection dominated by a single speaker. The described application that is implemented using supervised adaptation methods, is the first step in developing an unsupervised adaptable system. First, we analysed the speech recognition performance on this collection using a standard broadcast news system that scores a 30% WER on a BN domain test set. Although at TREC-8, BN systems for American-English produced error rates below 20%, the figures can be considered an adequate baseline given that it concerns an unadapted, single-pass system trained on 22 hours of acoustic training data and with a standard 65K newspaper vocabulary and language model.

As expected, deploying the broadcast news system for transcribing oral history data resulted in high error rates of around 80% WER. In order to improve on the BN system, several adaptation schemes were applied on the acoustic level and on the language model level. Using a SMAPLR adapted acoustic model, a 30K vocabulary optimised for the domain and a mixture language model of a newspaper LM, a general conversational-speech LM and a domain specific LM, gave the best overall results: 66.9%, a 13.5% absolute improvement on the baseline BN system.

Although near perfect transcripts are not required for a successful application of spoken document retrieval, error rates in this range are clearly insufficient. As domain specific acoustic training data for Dutch are hardly available for the oral history domain and are costly to develop, the application of unsupervised acoustic adaptation methods will receive special attention. We also aim at training more robust general acoustic models using the Spoken Dutch Corpus. In addition, we are implementing an unsupervised data partitioning approach that segments speakers in a particular background condition, in order to apply acoustic model adaptation.

For further language model improvement we rely on our news corpus, complemented with transcripts of the Spoken Dutch Corpus. We aim at improving the broadcast news language models by applying language model adaptation schemes and by using higher order n-grams and class-based models. Domain specific text data in the oral history domain are hard to obtain. Although the number of OOVs that can directly be related to the WFH domain was relatively small, additional data can be helpful for optimising word selection, especially names. Using the internet as an additional text source has proven to be a useful strategy in the meeting domain [12].

Finally, we will investigate how our current speech recognition performance affects retrieval performance and query out-of-vocabulary rate. Users may use highly selective, domain specific-words in their queries that are infrequent in the ma-

terial. Because of their low frequency, such words have a low chance of being selected for the speech recognition vocabulary and thus become query out-of-vocabulary. Applying a word spotting or phone-lattice scanning approach for such OOVs would then be an option. In addition, the integration with NLP that is required for an adequate search environment based on speech transcripts, will be explored.

#### 5. Acknowledgements

This paper is based on research funded by the Dutch project MultimediaN. We like to thank the Willem Frederik Hermans institute for providing text and audio material.

#### 6. References

- [1] K. W. Church, "Speech and Language Processing: Where Have We Been and Where Are We Going?" in *Eurospeech-2003*, Genève, Switzerland, September 2003.
- [2] J. Goldman, et al., "Report of the EU/NSF working group on Spoken Word Audio Archives," <http://www.ercim.org/publication/ws-proceedings/Delos-NSF/SpokenWord.pdf>, 2003.
- [3] "ECHO Project Homepage," <http://pc-erato2.iei.pi.cnr.it/echo/>.
- [4] W. Byrne, D. Doermann, and M. Franz, "Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives," *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, July 2004.
- [5] "Willem Frederik Hermans Institute web-portal," <http://www.willemfrederikhermans.nl>.
- [6] J. Morang, R. Ordelman, F. de Jong, and A. van Hessen, "InfoLink: analysis of Dutch broadcast news and cross-media browsing," in *Proceedings of ICME 2005 (to appear)*, Amsterdam, September 2005.
- [7] R. Ordelman, "Dutch Speech Recognition in Multimedia Information Retrieval," Ph.D. dissertation, University of Twente, The Netherlands, October 2003.
- [8] I. Schuurman, M. Schoupe, H. Hoekstra, and T. van der Wouden, "CGN, an Annotated Corpus of Spoken Dutch," in *In Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, 2003.
- [9] B. Pellom and K. Hacioglu, "Recent Improvements in the CU Sonic ASR system for Noisy Speech: The SPINE Task," in *Proc. ICASSP*, 2003.
- [10] O. Siohan, T. Myrvoll, and C. Lee, "Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation," in *Computer, Speech and Language*, 16, 2002, pp. 5-24.
- [11] B. Raj, V. N. Parikh, and R. M. Stern, "The Effects Of Background Music On Speech Recognition Accuracy," in *Proc. of the ICASSP, Munich, Germany*, 1997.
- [12] T. Hain, et al., "The Development of the AMI System for the Transcription of Speech in Meetings," in *to appear: proceedings of MLMI2005*, 2005.