Rapid Development of an Afrikaans-English Speech-to-Speech Translator

Herman A. Engelbrecht

Department of E&E Engineering
University of Stellenbosch, South Africa
hebrecht@sun.ac.za

Tanja Schultz

Interactive Systems Laboratories
Carnegie Mellon University, USA
tanja@cs.cmu.edu

Abstract

In this paper we investigate the rapid deployment of a two-way Afrikaans to English Speech-to-Speech Translation system. We discuss the approaches and amount of work involved to port a system to a new language pair, i.e. the steps required to rapidly adapt ASR, MT and TTS component to AFrikaans under limited time and data constraints. The resulting system represent the first fully functional prototype built for Afrikaans to English speech translation.

1. Introduction

In this paper we describe the rapid deployment of a two-way Afrikaans to English Speech-to-Speech Translation system. This research was performed as part of a collaboration between the University of Stellenbosch and Carnegie Mellon University. Using speech and text data supplied by the University of Stellenbosch, a native Afrikaans speaker developed the Afrikaans automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS) components over a period of 2.5 months. The components were built using existing software tools created by the Interactive Systems Laboratories (ISL). The prototype is designed to run on a laptop or desktop computer using a close-talking headset microphone.

Afrikaans is a Dutch derivative that is one the 11 official languages in the Republic of South Africa. The 11 languages consists of 2 Germanic languages: English and Afrikaans, and 9 Ntu (or Bantu) languages: isiNdebele, Sepedi, SeSotho, Swazi, Xitsonga, Setswana, Tshivenda, isiXhosa, isiZulu. The majority of the population speaks two of the 11 languages: their native mother-tongue and English most often chosen as the second language. Therefore English can be regarded as the pivot language in South African culture and is the most natural choice to translate to and from. Afrikaans was chosen because of the following three reasons: (i) Of the remaining 10 official languages, Afrikaans has the longest written history and therefore the most available text data. (ii) Unlike the Ntu languages, Afrikaans has the same language root as English and therefore the similarities should help in developing Afrikaans-English translation. (iii) The developer is fluent in both Afrikaans and English, but does not speak any of the Ntu languages.

The paper is organised into four parts. In the first part we will discuss some of the characteristics of Afrikaans. In the second part we will present the system architecture of the prototype as well as discussing the different development strategies that were chosen for each component of the system. The third part will discuss the Afrikaans data resources that were available and the last part will discuss the implementation details and performance of the prototype system.

2. Language Characteristics of Afrikaans

The following discussion of the characteristics of Afrikaans has been obtained from [1].

2.1. History

Afrikaans is linguistically closely related to 17th century Dutch, and to modern Dutch by extension. Dutch and Afrikaans are mutually understandable. Other less closely related languages include the Low Saxon spoken in northern Germany and the Netherlands, German, and English. Cape Dutch vocabulary diverged from the Dutch vocabulary spoken in the Netherlands over time as Cape Dutch was influenced by European languages (Portuguese, French and English), East Indian languages (Indonesian languages and Malay), and native African languages (isiXhosa and Khoi and San dialects). The first Afrikaans grammars and dictionaries were published in 1875.

Besides vocabulary, the most striking difference from Dutch is the much more regular grammar of Afrikaans, which is likely the result of mutual interference with one or more Creole languages based on the Dutch language spoken by the relatively large number of non-Dutch speakers (Khoisan, Khoikhoi, German, French, Malay, and speakers of different African languages) during the formation period of the language in the second half of the 17th century.

2.2. Grammar

Grammatically, Afrikaans is very analytic. Compared to most other Indo-European languages, verb paradigms in Afrikaans are relatively simple. With a few exceptions, there is no distinction for example between the infinitive and present forms of verbs. Unlike most other Indo-European

Consonants	p b t tʃ d dʒ k g ʔ m n ɲ ŋ r ʀ f v w θ s ʃ z ʒ ɦ j l
Short vowels	i y u e ø ɛ œ ɔ a ə æ
Long vowels	iː yː uː eː øː oː ɛː œː ɜː ɔː aː æː
Diphthongs	iu ia ui eu oi ɔi ai aʊ aːi əi əu æy

Table 1: *Afrikaans phone set (IPA).*

languages, verbs do not conjugate differently depending on the subject e.g. "*ek is, jy is, hy is, ons is*" = Eng. "I am, you are, he is, we are".

Unlike in Dutch, Afrikaans nouns do not have grammatical gender, but there is a distinction between the singular and plural forms of nouns. The most common plural marker is the suffix -e, but several common nouns form their plural instead by adding a final -s. No grammatical case distinction exists for nouns, adjectives and articles, with the universal definite article being "*die*" = Eng. "the" and the universal indefinite article being " *'n* " = Eng. "a/an".

Vestiges of case distinction remain for certain personal pronouns. No case distinction is made though for the plural forms of personal pronouns, i.e "*ons*" means both "we" and "us"; "*julle*" means "you", and "*hulle*" means both "they" and "them". There is often no distinction either between objective pronouns and possessive pronouns when used before nouns.

In terms of syntax, word order in Afrikaans follows broadly the same rules as in Dutch. A particular feature of Afrikaans is its use of the double negative, something that is absent from the other West Germanic standard languages, e.g: "*Hy kan nie Afrikaans praat nie*" = Eng. "He cannot Afrikaans speak not" (literally). It is assumed that either French of San are the origins for double negation in Afrikaans. The double negative construction has been fully grammaticalized in standard Afrikaans and its proper use follows a set of fairly complex rules

2.3. Orthography

Written Afrikaans differs from Dutch in that the spelling reflects a phonetically simplified language, and so many consonants are dropped. The spelling is also considerably more phonetical than Dutch. Notable features include the use of 's' instead of 'z', hence South Africa in Afrikaans is written as "Suid-Afrika", whereas in Dutch it is "Zuid-Afrika". The Dutch letter combination 'ij' is written as 'y', except where it replaces the Dutch suffix -lijk, as in "*waarskynlik*" = Dutch "waarschijnlijk". The letters 'c', 'q' and 'x' are rarely seen in Afrikaans, and words containing them are almost exclusively borrowings from English, Greek or Latin. This is usually because words with 'c' or 'ch' in Dutch are transliterated as 'k' or 'g' in Afrikaans. The following special letters are used in Afrikaans: è, é, ê, ë, î, ï, ô û.

2.4. Phone Set

The Afrikaans phoneme set (shown in Table 1) consists of 27 consonants, 23 vowels and 12 diphthongs for a total of 62 phones. Vowels are further subdivided into 11 short vowels and 12 long vowels.

3. System Architecture

The target platform of the Afrikaans-English speech translation prototype is a desktop or laptop. Speech input is obtained using a standard PC sound card and a close-talking PC headset microphone. The demonstration prototype consists of 3 main components: ASR, MT and TTS. Each component was developed separately and then integrated into the prototype. The breakdown of the prototype system is shown in Fig. 1. The working of the speech translation prototype is broken into three actions:

1. Conversion of source language speech into source language text (ASR).
2. Translation of source language text into target language text (MT).
3. Conversion of target language text into target language speech (TTS).

The choices of the recognition, translation and synthesis strategies were heavily influenced by the amount of labor-intensive work and time that is required to implement each strategy. Data-driven techniques were preferred over knowledge-based techniques as it would enable the prototype to be developed more rapidly. The following strategies were therefore chosen:

- For the speech recognition a statistical n-gram language model based recognition strategy was chosen as this does not involve the labor-intensive task of writing recognition grammars.
- For the translation strategy a statistical machine translation (SMT) approach was chosen instead of an Interlingua based approach. An Interlingua based approach would require the development of a part-of-speech tagger, an analysis grammar and a generation grammar. The SMT approach only requires the development of a translation model (TM) and a statistical language model (SLM), both which can be learned directly from text data.
- For the synthesis strategy a concatenative speech synthesis approach was chosen as a first implementation. Concatenative speech synthesis requires the construction of databases of natural speech for the target domain. A new utterance in the target domain is synthesized by selection and concatenation of appropriate subword units. The disadvantage of unit-selection concatenative speech synthesis is that it requires large amounts of memory.

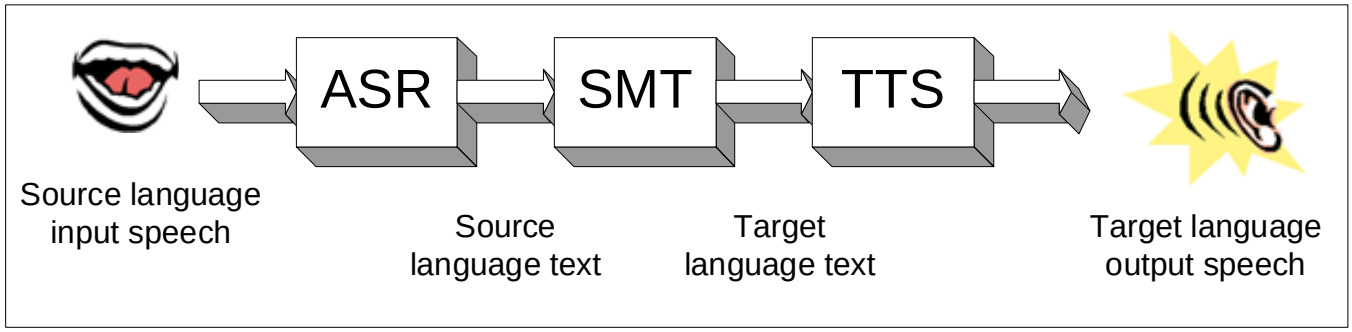

Figure 1: *The system architecture of the Afrikaans-English speech translation prototype.*

For each of the main components it was necessary to develop the following subcomponents:

- ASR: Acoustic Models, Language Models and Pronunciation Dictionary.
- SMT: Translation Models and Language Models.
- TTS: Pronunciation Dictionary and Letter-To-Sound Rules.

The main components were finally integrated by simply using the output of each preceding component as the input of the next component. The best ASR output was used as input for the SMT component and the best SMT translation output was used as input for the TTS component. Only the first best ASR output was used as input for the SMT component. No effort was made to compensate for recognition errors (by using word lattices as input) or for speech disfluencies. that are sometimes used in an attempt to reduce the impact of using recognised speech as input instead of text, on SMT performance.

4. Language Data Resources

The biggest challenge to developing the system was the limited amount of available Afrikaans speech and text data. Over the past 100 years Afrikaans has developed a rich literature which results in the accumulation of large text data. in contrast, very little efforts have been undertaken so far to record and transcribe spoken speech (suitable for speech recognition). In order to develop the translation component, it is necessary to use parallel text data. The text data is required for the development of the statistical language models needed for both the ASR and SMT components. It is also required for the development of the translation models (TM) needed for the SMT component. Parallel text data is more difficult to create and only 43k utterances could be obtained. Acoustic model (AM) development require transcribed speech data. In total there was only about 6 hours of transcribed Afrikaans speech data available. Furthermore, the transcribed speech data was recorded over landline and cell phone network. As the prototype was designed to be used with a close-talking PC headset microphone, a channel mismatch would have occurred if only the available Afrikaans speech was used for training the acoustic models. In order to reduce the channel mismatch it was decided to collect a limited amount of Afrikaans speech under the same acoustic conditions as the target application. In the rest of this section we will describe the data resources in more detail.

4.1. Text Data

The text data consists of multilingual parliament sessions that were translated into both Afrikaans and English. The data consists of 39 parliamentary sessions from the year 2000-2001 for a total of 43k parallel sentences. The sentence lengths are distributed from sentences that are single words to sentences that are more than 100 words long. The translated parliamentary sessions are commonly referred to as Hansards. In the rest of the paper we will refer to the parliamentary domain as the Hansard domain.

4.2. Speech Data

4.2.1. AST data

The Afrikaans speech data was collected during a period of 3 years ending in March 2004 by a consortium known as African Speech Technology (AST) [2, 3]. The AST speech corpus consists of 5 languages for a total of 11 dialects. The data was collected over the telephone and cellphone networks and each participant had to read a datasheet containing 40 utterances. This included a phonetically balanced sentence consisting of 40 words for each dialect. The transcriptions of the AST data are orthographically and phonetically transcribed. Speech and non-speech utterances have also been marked and the phonetic transcriptions have been corrected by hand. Only the mother-tongue Afrikaans speech data was used in this research (referred to as the AA data). The AA speech data consists of a total of 265 speakers, 113 male and 152 female, for a total of 10768 utterances. 191 of the recordings were made using landlines and 74 of the recordings were made using the cell phone network.

4.2.2. Hansard data

In order to be able to evaluate the complete demonstration prototype (excluding the synthesis) it was necessary to record utterances that are representative of the Hansard domain. As there was only two native Afrikaans speakers, it was decided to record 1,000 utterances (500 utterances per speaker). The utterances were recorded at a sampling frequency of 16kHz using a laptop and a close-talking PC headset microphone (Andrea Anti-noise NC-61). The utterances were recorded in a medium-sized room with low to medium noise levels. The 1,000 sentences were chosen from the parallel text data so that the distribution of sentence lengths in the evaluation data would be representative of the distribution found in the parallel text corpus (up to a sentence length of 40 words per utterance). The utterances are classified as read speech, as the utterances were recorded by prompting the speaker. The utterances were only orthographically transcribed and no manual time-alignment of the speech signal and transcription were performed.

4.2.3. Pronunciation Dictionaries

As the AST speech data had been ortographically and phonetically aligned, a pronunciation dictionary containing 5,361 words can be extracted from the transcriptions. The AST pronunciation dictionary has a vocabulary size of 3,795 words and a total of 1.41 pronunciation variants (rounded to the second decimal). Another syllable annotated pronunciation dictionary, developed by the University of Stellenbosch, was also available. The Stellenbosch dictionary has a vocabulary size of 36,783 words and does not contain any pronunciation variants. By combining the AST dictionary and the Stellenbosch dictionary a new dictionary was formed that has a vocabulary size of 38,960 words and a total of 1.08 pronunciation variants.

5. Development of System Components

5.1. Partitioning of data sets

In order to be able to evalaute the complete prototype as well as each component separately, it was decided to use the same evaluation set for all evaluations. As previously mentioned 1,000 utterances were selected from the parallel text data and recorded using a close-talking microphone. The 16kHz Hansard utterances are downsampled to 8kHz in order to match the acoustic models. The 200 longest utterances were used for adaptation of the recogniser and the remaining 800 utterances were used for evaluation purposes (which will be referred to as the Hansard evaluation set). The rest of the 41k sentences were used for the development of the translation models. In Table 3 information regarding the Afrikaans and English parallel text data is shown. Although the Afrikaans text data only has a vocubulary size of 25k words and the pronunciation dictionary consists of 39k words, not all the words in the Afrikaans text data were covered by the pronunciation dictionary. The following three constraints were used when selecting the 1,000 sentences to be recorded:

1. Every word in a recorded sentence had to be covered by the pronunciation dictionary.

2. The distribution of words per sentence had to be representative of the distribution in the training data.

3. No sentence containing more than 40 words were recorded.

The AST speech data was divided into training, development and evaluation sets which each respectively consists of 70%, 15% and 15% of the AST data. The AST training data contains 187 speakers and 7696 utterances.

5.2. Automatic Speech Recognition

The Afrikaans acoustic models were bootstrapped from the GlobalPhone [4, 5] MM7 multilingual acoustic models using a web-based tool called SPICE [6]. The MM7 phones did not cover all the Afrikaans phones and it was decided to reduce the 62 phone set to 39 phones which was done by splitting the diphthongs into two separate phones and by not distinguishing between long and short vowels. It is unknown what the impact of the large reduction in the phone set has on the ASR performance. Another possibility would have been to bootstrap unknown Afrikaans phones with neighboring phones, but unfortunately time did permit the development of a Afrikaans system with a larger phone set. CMU's Janus JrTk [7, 8] was used to train the acoustic models on 4.2 hours of the AST speech data.

As the recogniser will be used with a close-talking headset microphone a channel mismatch exists between the evaluation conditions and the training conditions. There also exists a domain mismatch as the AST data covers various tasks (as described in section 4.2.1) while the Hansard data covers parliamentary debates. In an attempt to adapt to the acoustic environment and the domain, the acoustic models are further trained on 200 utterances of Hansard speech data. The acoustic models were adapted by simply training on the Hansard speech data and not by using MLLR or MAP adaptation. However, as the Hansard speech data consists of only two speakers, this further training probably adapted to the test speakers rather than the evaluation conditions. The Afrikaans recogniser is a fully-continuous 3-state HMM recogniser with 500 triphone models (tied using decision trees). Each state consists of a mixture of 128 Gaussians. The frontend uses 13 MFCCs, power, and the first and second time derivatives of the features. These are reduced to 32 dimensional feature vectors using LDA. Both vocal tract length normalisation (VTLN) and constrained MLLR speaker adaptive training (SAT) was employed when training.

The Afrikaans and English language models were trained using SRI's statistical language toolkit SRILM [9]. The Afrikaans LM is a trigram language model with a perplexity

of 103.71 and a OOV rate of 0.0% on the Hansard evaluation set. It was trained on 694,455 words and a vocabulary of 25,623 words.

Both the Hansard adapted acoustic models and the unadapted acoustic models were evaluated on the Hansard evaluation set which consists of 15,259 words and has a vocabulary size of 2.45k words. The results are shown in Table 2. It can be seen that the unadapted acoustic models has a fairly poor performance of 46.5% WER. Fortunately the acoustic models that were adapted to the Hansard evaluation conditions has a WER of only 20.0% which is a relative improvement of 54.3%. Thus the channel and domain mismatch that exists between the training conditions and the evaluation conditions are partially solved by adapting on the Hansard data. The speaker-independency of the Afrikaans recogniser could not be determined (as a result of the limited number of available Afrikaans speakers), but because the Hansard adaptation data only contains two native Afrikaans speakers the Afrikaans recogniser is quite possibly very speaker-dependent. It can also be seen that the ASR performs significantly better for the male speaker than for the female speaker.

	Unadapted AMs	Adapted AMs
Number of words	15,259	15,259
Vocabulary size	2,450	2,450
Pronunciation variants	1.08	1.08
OOV	0.0%	0.0%
Trigram LM PP	103.71	103.71
WER (male)	39.1%	17.6%
WER (female)	54.0%	22.3%
WER (total)	46.5%	20.0%

Table 2: *ASR evaluation results on the Hansard set.*

The total development time for the ASR component is estimated to be 8 weeks and was the most difficult and time-consuming component to develop.

5.3. Statistical Machine Translation

According to [10] statistical machine translation defines the task of translating a source language sentence ($\mathbf{f} = f_1 \dots f_J$) into a translation sentence ($\mathbf{e} = e_1 \dots e_I$) of the target language. The SMT approach is based on Bayes' decision rule and the noisy channel approach in that the best translation sentence is given by:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} [P(\mathbf{e}|\mathbf{f})] = \arg\max_{\mathbf{e}} [P(\mathbf{f}|\mathbf{e})P(\mathbf{e})] \quad (1)$$

where $P(\mathbf{e})$ is the language model of the target language and $P(\mathbf{f}|\mathbf{e})$ is the translation model. The $\arg\max$ denotes the search algorithm, which finds the best target sentence given the language and translation models. For a detailed discussion of CMU's statistical machine translation system refer to [11]. The system contains a IBM1 lexical transducer, a phrase transducer and a class based transducer. Only the IBM1 lexical transducer, which is a one-to-one lexicon mapper, is used in this research. The language model is n-gram based and up to trigrams are used. The decoder is a beam search based on dynamic programming combined with pruning.

As the intention was to develop a two-way speech translation demonstration prototype, both Afrikaans and English translation systems were developed. The translation models were trained on the 42k Hansard parallel data and was evaluated using the same 800 Hansard sentences that were used to evaluate the ASR component. The same Afrikaans SLM was used as was trained for the ASR component. The English SLM is also a trigram language model with a perplexity of 86.62 and a OOV rate of 0.0% on the Hansard evaluation set. It was trained on 687,154 words and a vocabulary of 17,898 words.

The influence of punctuation on SMT performance was investigated. In the first case all punctuation was removed from the parallel text before training and in the second case the punctuation was left in the data. Separate SLMs were also trained for the systems with and without punctuation. Table 3 summarizes the information regarding the Afrikaans and English text data. It is interesting to note that the Afrikaans vocabulary size is 43% larger than English vocabulary size. Although Afrikaans is much less inflected than English, Afrikaans has less rigid spelling rules regarding the formation of compound words. Afrikaans compound words can be written in three different ways: (i) as a single word, (ii) as separate words or (iii) as separate words connected with dashes. When preparing the text data, no effort was made to force the Afrikaans text to conform to a single method of forming compound words. It has also been noticed that Hansard domain contains a large number of compound words which results in the large vocabulary size for Afrikaans.

Text Data Language	English	Afrikaans
Number of Sentences	41,239	
Number of Words	687,154	694,455
Vocabulary Size	17,898	25,623
LM Perplexity w/o punct.	87.21	103.71
LM Perplexity with punct.	62.28	72.28
OOV in Testset	0.0%	0.0%

Table 3: *Parallel Corpus Statistics.*

In Table 4 the results of the SMT experiments are shown for both Afrikaans-English and English-Afrikaans translation. It can be seen that Afrikaans-English translation does benefit from the use of punctuation as both the NIST and the BLEU metric increase slightly. For English-Afrikaans translation the NIST metric is degraded slightly by the use of punctuation although the BLEU metric is increased. This would seem to indicate that the fluency of the translation

benefits from punctuation although the accuracy is not significantly affected. Both the Afrikaans-English translation and English-Afrikaans translation results are comparable to the IBM1 lexical transducer results (BLEU 41.6, NIST 7.58) that were obtained on the Chinese-English parallel corpus of 20,000 sentences [12]. There is still much scope for improvement as only the most simple of translation models was applied. The total development time for the SMT component

	Afrikaans-English		English-Afrikaans	
Results	BLEU	NIST	BLEU	NIST
IBM1 w/o punct	34.13	7.65	34.68	7.93
IBM1 with punct	36.11	7.66	34.81	7.73

Table 4: *SMT evaluation results on the Hansard test.*

is estimated to be 1 week and was relatively easy when compared to the ASR development.

5.4. Speech Synthesis

A limited domain Afrikaans voice is built using the Festival Speech Synthesis System [13]. A male Afrikaans unit-selection voice was built following the techniques for building synthetic voices in new languages developed by CMU [14, 15, 16]. The same phone set is used for synthesis as was used for the recogniser. The 500 Hansard utterances that was used for adaptation and evaluation of the recogniser were used for building the unit-selection voice. We were also fortunate to obtain a syllable annotated pronunciation lexicon of 36,783 Afrikaans words. It was therefore not necessary to build a pronunciation lexicon for Afrikaans.

A statistical letter-to-sound rule model was trained on 90% of the pronunciation dictionary and evaluated on the remaining 10% [17]. The evaluation pronunciations were chosen by selecting every 10th word in the alphabetically sorted pronunciation dictionary. The results of the letter-to-sound rules are shown in Table 5. The letter-to-sound rules managed to correctly predict 85.24% of the words which is to be expected as Afrikaans spelling reflects a phonetically simplified language. These results are comparable to the results German (89.38% word correct) [17].

Trainset pronunciations	33,121
Testset pronunciations	3,680
Phones correct	97.92%
Words correct	85.24%

Table 5: *Evaluation of Letter-To-Sound rules.*

As only two Afrikaans speakers were available it was not possible to formally evaluate the performance and quality of the Afrikaans speech synthesis. In all cases the Afrikaans pronunciations were understandable, but the following informal observation can be made regarding the quality of the synthesis:

- The Afrikaans phone set made no distinction between long and short versions of the same vowel. Consequently some pronunciation errors were made when words contained long vowels.
- The lack of diphthongs in the phone set resulted in some incorrect pronunciation of words containing diphthongs.

Both of these problems can be corrected by simply using a larger phone set which includes the diphthongs and models both long and short vowels.

The total development time of the synthesis component is estimated to have been one week. The availability of a 37k Afrikaans pronunciation dictionary shortened the development of the synthesis component considerably.

5.5. Estimate of total development time

Table 7 summaries the estimate of the total system development time.

	Time	
Task	Days	Weeks
Familiarise with ASR software	7	1.4
Preparation of data for AMs	5	1
Phone set adaptation of dictionary and transcriptions	1	0.2
Bootstrap & Training of AMs	8	1.6
Tuning of ASR and adaptation of AMs	7	1.4
Preparation of data for LMs	11	2.2
Generation of LMs	1	0.2
Preparation of data for LMs	3	0.6
Generation of TMs	2	0.4
Preparation of data for TTS	1	0.2
Generation of Afrikaans voice	1	0.2
Generation of LTS rules	3	0.6
Familiarise with 'one4all' framework	2	0.4
Integration of components	3	0.6
Evaluation	5	1
Total development time	60	12

Table 7: *Estimate of system development time.*

6. Prototype Translation System

6.1. Description

For the development of the prototype we used the "one4all demonstrator systm platform"' as described in [18] and essentially the same software framework was used. The following was done to develop the prototype: (i) the recogniser was replaced with an Afrikaans recogniser; (ii) the SMT transducers were replaced with Afrikaans-English and English-Afrikaans transducers; and (iii) the speech synthesis voice was replaced with an Afrikaans voice.

Results	WER	Afrikaans-English			
		NIST		BLEU	
		SCORE	Rel. Improvement	SCORE	Rel. Improvement
TEXT w/o punct	0.0%	7.65	-	34.13	-
ASR w/o punct	20.0%	6.12	-20.0%	25.45	-25.4%
ASR w/o punct	46.5%	4.56	-40.4%	17.39	-49.0%
TEXT with punct	0.0%	7.66	-	36.11	
ASR with punct	20.0%	6.04	-21.1%	24.42	-32.4%
ASR with punct	46.5%	4.40	-42.6%	16.72	-53.7%

Table 6: *Prototype evaluation results.*

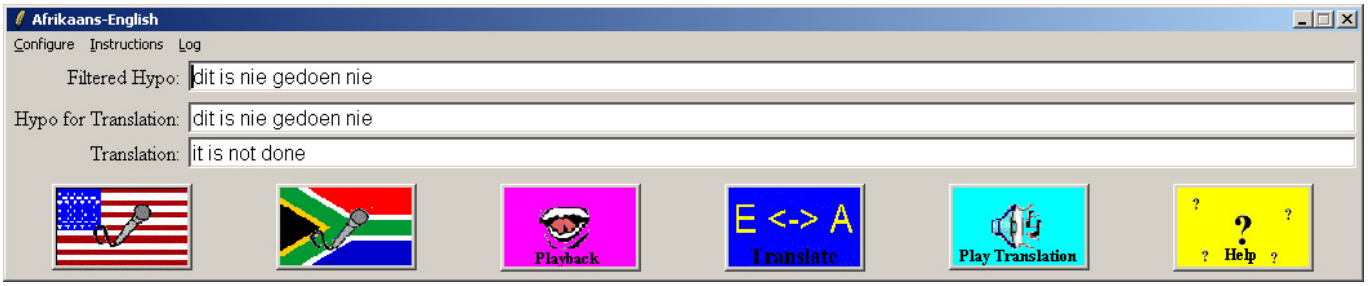

Figure 2: *An example of Afrikaans-English translation prototype.*

The integration, adaptation and evaluation of the prototype system is estimated to have taken one week. Figure 2 shows the interface of the demonstration prototype system.

6.2. Evaluation

The complete prototype was evaluated in order to determine the influence of imperfect recogniser on the translation. Only the Afrikaans-English speech-to-speech translation was evaluated by using the single best recognition result of the recogniser as input to the SMT engine. The results are shown in Table 6. The best result of 6.12 on the NIST metric and 25.45 on the BLUE metric is obtained when not using punctuation. As expected the translation performance of the best results is significantly affected as the translation accuracy drops by 20.0% relative and the fluency of the translation drops by 25.4% (as respectively measure by the NIST and BLEU metric). Overall the use of punctuation results in worse translation performance than not using punctuation. This is to be expected as the ASR component is does not add punctuation to the recognition output. It seems that there is a correlation between the WER of the recogniser and the degree by which the translation accuracy is affected, but further experiments are required in order to confirm this theory.

7. Conclusion

In this paper we presented the rapid 2.5-month development of an Afrikaans-English speech-to-speech translation demonstration system. The recognition component is still the most challenging component to develop as can be seen by the 20% word-error-rate performance of the Afrikaans recogniser. The Afrikaans-English translation results of 36.11 on the BLEU metric and 7.66 on the NIST metric is very encouraging as it is comparable to the results obtained on other language pairs such as Chinese-English (when using similar translation modelling and a similar size parallel data corpus). As only the most simple statistical translation models were used there is much scope for improvement. Lastly, the evaluation of the complete demonstration prototype shows that errors in the recognition output degrades the translation results, as expected. There seems to be a correlation between the WER of the recogniser and the accuracy of the translation (as measured by the NIST metric) but further experiments are required to confirm this theory.

8. Acknowledgement

The authors wish to thank the following persons for their contributions: Paisarn Charoenpornsawat, Alan Black, Matthias Eck, Bing Zhao, Szu-Chen Jou, Susanne Burger and Thomas Schaaf.

9. References

[1] Wikipedia, “Afrikaans — Wikipedia, the free encyclopedia,” 2005, [Online; accessed 27-June-2005]. [Online]. Available: http://en.wikipedia.org/wiki/Afrikaans

[2] Roux, J.C, Botha, E.C. and Du Preez, J.A., “Developing a Multilingual Telephone Based Information System in African Languages,” in *Proc. of 2nd Intl.*

Language Resources and Evaluation Conf., Athens, Greece, June 2000.

[3] Roux, J.C., "Final Report on the African Speech Technology (AST) Project," University of Stellenbosch, Tech. Rep., Feb. 2005.

[4] Schultz, T., "GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University," *Proc. of the Int. Conf. of Spoken Language Processing*, Sept. 2002.

[5] Schultz, T., Waibel, A., "Language-independent and language adaptive acoustic modelling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[6] Schultz, T., "Towards Rapid Language Portability of Speech Processing Systems," in *Conference on Speech and Language Systems for Human Communication*, Delhi, India, Nov. 2004.

[7] Finke, M., Geutner, P., Hild, H., Kemp, T. Ries, K. and Westphal, M., "The Karlsruhe Verbmobil Speech Recogntion Engine," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Munich, Germany, 1997.

[8] Soltau, H., Metze, F., Fügen, C. and Waibel, A., "A one-pass decoder based on polymorphic linguistic context assignment," in *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, 2001.

[9] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit," in *Proc. of ICSLP*, Denver, Colorado, Sept. 2002.

[10] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L., "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, 1993.

[11] Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B. and Waibel, A., "The CMU Statistical Machine Translation System," in *Proc. of the MT Summit IX*, New Orleans, USA, Sept. 2003.

[12] Vogel, A., Hewavitharana, S., Kolss, M., and Waibel, A., "The ISL Statistical Translation System for Spoken Language Translation," in *Proc. of IWSLT*, Kyoto, Japan, 2004.

[13] Black, A., Taylor, P. and Caley, R., "The Festival Speech Synthesis System," 1999. [Online]. Available: http://festvox.org/festival

[14] Black, A. and Lenzo, K., "Building Voices in the Festival Speech Synthesis System," 2000. [Online]. Available: http://festvox.org/bsv

[15] Black, A. and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. of Eurospeech*, vol. 2, Rhodes, Greece, 1997.

[16] Black, A. and Lenzo, K., "Limited Domain Synthesis," in *Proc. of ICSLP*, vol. 2, Beijing, China, 2000.

[17] Black, A., Lenzo, K. and Pagel, V., "Issues in Building General Letter to Sound Rules," in *Proc. of 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[18] Suebvisai, S., Charoenpornsawat, P., Black, A., Woszczyna, M., and Schultz, T., "Thai Automatic Speech Recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, 2005.