Linguistics and HLT for Countries with Minimal ICT Infrastructure

Lerato Lerato, Matšepo Machobane and *Mope Lephoto*
National University of Lesotho
P. O. Roma, 180
Lesotho

Abstract

This paper presents a visionary Human Language Technologies (HLT) model aimed at enabling the Basotho, who are computer illiterate, and thus unable to access internet but have access to either a landline handset or a cellular phone, to browse and access electronic government information service in their language, Sesotho. This requires various speech-based HLT components such as pronunciation modelling, speech recognition, text-to-speech (TTS) and language identification. For purposes of this presentation, we concentrate on speech recognition and TTS. The paper also provides a preliminary experimental setup necessary for the proposed model. The proposed model is specifically intended to empower the illiterate and the underprivileged members of the society.

Introduction

Information and Communications Technology (ICT) is a catch-all phrase encompassing electronic information-processing technologies such as computers and internet, fixed-line telecommunications, mobile phones and wireless communications such as radio and television broadcast. ICT is considered one of the most important fields of research in both developed and the developing countries as it can promote innovation, increase productivity and enrich the quality of lives (Riley, 2005). Despite the fast growing ICT in the developed countries, most African countries still have low electricity coverage and poor telephony infrastructure. ICT, in the form of mobile phones, has taken relatively large strides compared to other forms of communication in Lesotho. Table in 1 shows figures which summarise ICT

coverage countrywide according to Lesotho Telecommunication Authority (LTA) records (http://www.lta.org.ls).

Table 1: ICT coverage in Lesotho

Mode of Communication	National Coverage (%)
Internet	Less than 1
Mobile	11.6
Landline	2.1
Radio Broadcast	75

According to Table 1, less than 1 percent Basotho have access to the Internet, 11.6 percent own mobile phones, 2.1 percent have landline telephone connections and 75 percent have access to radio broadcast. We set aside the radio broadcast, as it can only be used on a limited scale to access information. One can speculate that every employed Mosotho with tertiary qualification owns either a mobile phone and/or a landline telephone. It is highly probable that a high percentage of people who own mobile phones do not know English. Due to limited computer knowledge and English communication skills many Basotho are unable to undertake transactions with ICT-based services such as the Lesotho government website. This means that the Lesotho government website is only accessible to the extreme minority of the nation. According to the chief executive officer for Commonwealth Communications Organization (Spio-Garbrah, 2005), 85 percent of the world's poor people live in villages and 80 percent of them have never used the internet. This paper proposes a model that shows how language and speech technology can be used by computer illiterate people in poor countries such as Lesotho. Therefore, the main objectives of the paper are to:
- describe HLT components relevant for this study;
- describe linguistic engineering in HLT;

- propose the visionary HLT-ICT model which demonstrates how HLT application utilises ICT infrastructure;
- describe experiments and relevant tests undertaken;
- provide some conclusions and indications of how the proposed model empowers computer illiterate society.

There are related, but not identical, on-going studies on the telephone-based delivery of government services from which our study benefited (Barnard, Cloete and Patel, 2004; Naidoo, Barnard and Davel, 2005).

The paper is organised as follows: Section one provides a brief description of HLT and its components relevant to the study. It further describes how speech recognition and Text-To-Speech (TTS) fit into this field. Section two describes the linguistics engineering in HLT. Section three presents the visionary model, which demonstrates how HLT application can utilise ICT infrastructure. Section four describes the relevant experimental setup and methodology for the proposed model. Section five presents accomplished tests and experiments.   Finally, section six provides some conclusions and recommendations for the way forward.

1.   Human Language Technologies

HLT is the term used to describe the language capabilities designed into the computing applications used in information technology systems. It has attracted a lot of attention in the research done by Linguists, Computer Scientists and Engineers (Dutoit, 1997; T-Systems, 2004). The HLT requires language engineering for most of its components. It is divided into text-based and speech-based technologies.   The text-based technologies include topics such as language processing, spell checking, lexicography, semantics, morphological and syntactic analysis.   This is where computational models of linguistic phenomena are

created in order to enable, for example, natural language processing and text editing. We provide a brief description of the speech-based technologies such as speech recognition and text-to-speech, which are used as examples of HLT in the illustration of how Linguistics and ICT integrate.

1.1   Speech recognition

Speech recognition as an independent technology is any system that acquires voice and converts it to text that can be read on the computer screen or on the mobile phone as short message services (SMS). The input to speech recognition system is the speech that goes through the microphone, telephone handset or cellular phone such that it is digitised by the time it gets to the computer. This digitised speech is converted to text that can be used as query information to some database storage or displayed on the computer monitor (Rabiner and Juang, 1993).

One of the modules of speech recognition system is its language pack. The language pack includes language model, grammar and dictionary. A language model contains the characteristics of the dynamics of language of interest in the implementation, Sesotho in our case. Grammar acts as a constraint of channelling each user to only use certain words or phrases that are relevant to the application. The dictionary contains all the words related to the application so that the system is able to correlate pronunciation with the acoustic information concerned with measuring and analysing the physical properties of sound waves produced when we speak. Figure 1 reflects the use of language components in the speech recognition process as implemented in the HTK software (Rabiner 1989). The components of this system are described in sections 1.1.1 and 1.1.2 below.
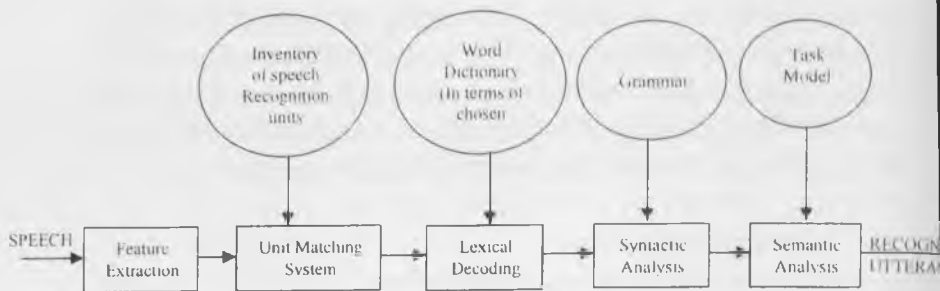
Figure 1: The generic speech recognition system

1.1.1   *Feature Analysis and Unit Matching System (Rabiner 1989)*

Feature extraction is the component, which extracts information related to a sound of a word, phrase or sentence. This input word, phrase or sentence is technically called *speech*. The speech is compared (pattern matched) with the stored speech. Signal processing and classification techniques are used in the matching of stored speech and input speech. The stored speech in the *inventory* (Owens, 1993) can exist in various ways. Possible inventory units, which a speech recognition system designer can choose, include linguistically based sub-word units such as phones, diphones, demisyllables, syllables, fenones and even acoustic units. In addition, word units can be used to store speech in the inventory. Most Hidden Markov Model (HMM) (http://htk.eng.cam.ac.uk ; Rabiner, 1989) based speech recognition such as the one used in the HTK system largely use phones as storage units. The word dictionary or lexicon contains all the words that the recogniser can possibly use. For example, the phrase, *Ke rata Sesotho* 'I like Sesotho' can be stored as, *"ki:_rA:tV_s@su:tU"* in the dictionary. This phrase can be lexically decoded so that the processing can incorporate linguistic decoding and computer language during recognition.

1.1.2 Syntactic and Semantic Analyses

The syntactic analysis decides on the category that effectively corresponds to each linguistic unit, given the context of the sentence. This module as shown in figure 1, places constraints of pattern matching in that, if a word or phrase is outside the scope of the speech recognition application, then the recognition is unsuccessful. The *grammar* defines all the words that an application requires. Any word that is not contained in the *grammar* will not be recognised. Semantic analysis adds further constraints to the set of recognition search alternatives. Some syntactically correct words or sub-words can be eliminated from the system during the semantic analysis depending on the task of the speech recogniser, which is stored in the *task model*. If the input speech matches all the steps shown in figure 1, the recognised speech is converted into text (recognised utterance).

1.2 Text-to-Speech

Text-speech (TTS) is a type of speech synthesis application that is used to create a spoken version of the text in a computer document. A TTS technology enables people to type information that can be read out as speech. The TTS system has the ability to use the language dynamics that exist in the sent text and from that it can read text in one's chosen language. The tuning of the TTS system is also related to the language in which the services are offered. The TTS dictionary can also be customised to one's own language. Furthermore, the voice produced by the TTS engine can also be customised to local accents and languages. The TTS takes an utterance structure until a waveform that says what the text contains is built. First the text is divided into words, then the utterance is chunked into prosodic phrases and then pronunciation is obtained from a lexicon to sound rule system, including phonetic and syllable structure. Finally the intonation accents are assigned to appropriate syllables and then the waveform is rendered

is a generic functional diagram of the TTS system according to Dutoit (1997). In this system, the typed text undergoes natural language processing which includes the phonetic transcription of the text together with desired intonation (termed prosody). A Digital Signal Module transforms the symbolic information it receives into speech. The TTS process was implemented by the authors in the laboratory using the *festival* (Black, Taylor and Caley, 1999) TTS software. The implementation is illustrated in figure 3 where the Sesotho text, *Lerato ke mang* 'Who is Lerato?' was typed as the input to the TTS system, which processed it, and the resulting speech waveform (*"lerato ke mang"* in audio) is shown in the same diagram.
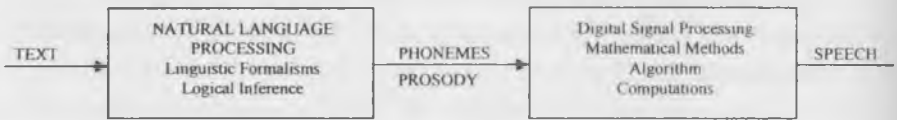
Figure 2: A Generic TTS system diagram (Dutoit, 1997)
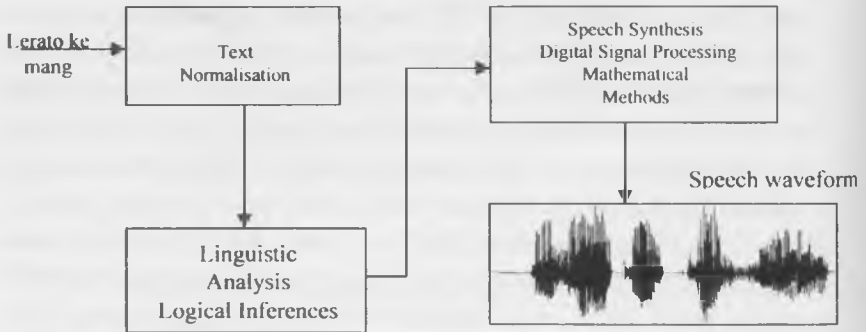
Figure 3: The experimental result of the in-house implementation of TTS

Prior to 2004 none of the HLT systems described in this section had been applied to African Languages

(http://www.meraka.org.za/hlt_projects.htm). As part of the Local Speech Technology initiative isiZulu TTS system (Louw, Davel and Barnard 2005) has been developed. Attempts are being made to develop systems in other languages. There are several other on-going projects and studies dealing with different aspects of HLT in different African languages (Tucker and Shalonova 2004; Louw, Davel and Bernard 2005). None of the HLT system developed for one language can suit another language, as languages may differ with respect to tone, stress, morphological, syntactic and morpho-syntactic characteristics.

2. Linguistics in HLT

Richards, Patt and Platt (1992:215) define the term linguistics as the study of language as a system of human communication. Linguistics comprises components such as phonetics (the study of inventory of speech sounds), the lexicon (the dictionary), syntax (the system of rules and categories that underlies sentence formation in human languages) and semantics (the study of meaning in human language). Linguists play a leading role in the implementation of a HLT application (T-Systems 2004; http://www.sita.co.za). Speech Scientists rely on the knowledge of the language when designing or tuning any speech technology application. The dynamics of building a viable language model require intensive linguistic engineering. Concentrating on the speech recognition and TTS applications, it is found that the dictionary, which the designers require, is created and optimised by the linguists. An example that some companies have used is the breaking down of words into phonetic units. These units are transcribed using the universal system for transcribing speech sounds known as the International Phonetic Alphabet (IPA). Several companies have converted the conventional IPA phonetic symbols into the computer readable phonetic notation

(http://www2.arts.gla.ac.uk/IPA/ipa.html). The computer uses ASCII code as a means of encoding any character that has been inserted through a keyboard. For this reason, there must be a way of transforming the IPA phonetic notation into what a machine can realise. One method of converting the standard IPA phonetic symbols into the machine-readable code is called SAMPA (speech assessment methods phonetic alphabet). The process of converting IPA to SAMPA is called phonetic transcription. This phonetic alphabet (SAMPA) was developed by the international group of phoneticians and was applied for the first time to various European Languages. Table 2 provides an example of SAMPA transcriptions.

**Table 2**: IPA phonetic symbols to SAMPA symbols

IPA Vowel Symbols	SAMPA
[æ]	{
[ʊ]	U
[ʌ]	V
[ə]	@
[ɜ]	3
[ɛ]	E
[ɪ]	I
[ɔ]	O

3. The Visionary HLT-ICT Integrated Model

According to the Lesotho Government website, in 2003 the estimated literacy rate in Lesotho was 85 percent (http://www.lesotho.gov.ls/articles/2005/). This suggests that if there was Internet everywhere in Lesotho, and that it

could be browsed in Sesotho, 85 percent of Basotho could readout information online. However, most of those internet users would have to be computer literate but it would be costly to implement computer centres and the Internet. This section focuses on how language and speech technology can empower those Basotho who have no Internet access but have a mobile phone or can, at least, have some point where they can access a telephone box.

The HLT can be incorporated into the ICT infrastructure so that users of the system can be able to *talk* to the system when accessing the same information as the Internet browsers. Figure 4 shows a possible model whereby HLT facilitates the *voice* browsing of government information. Subsection 3.1 describes the call flow from a mobile phone or the handset to the *Government Information Bank.* The normal web browsing process is also described in the same subsection. An example of how a typical application can be designed is shown in subsection 3.2.
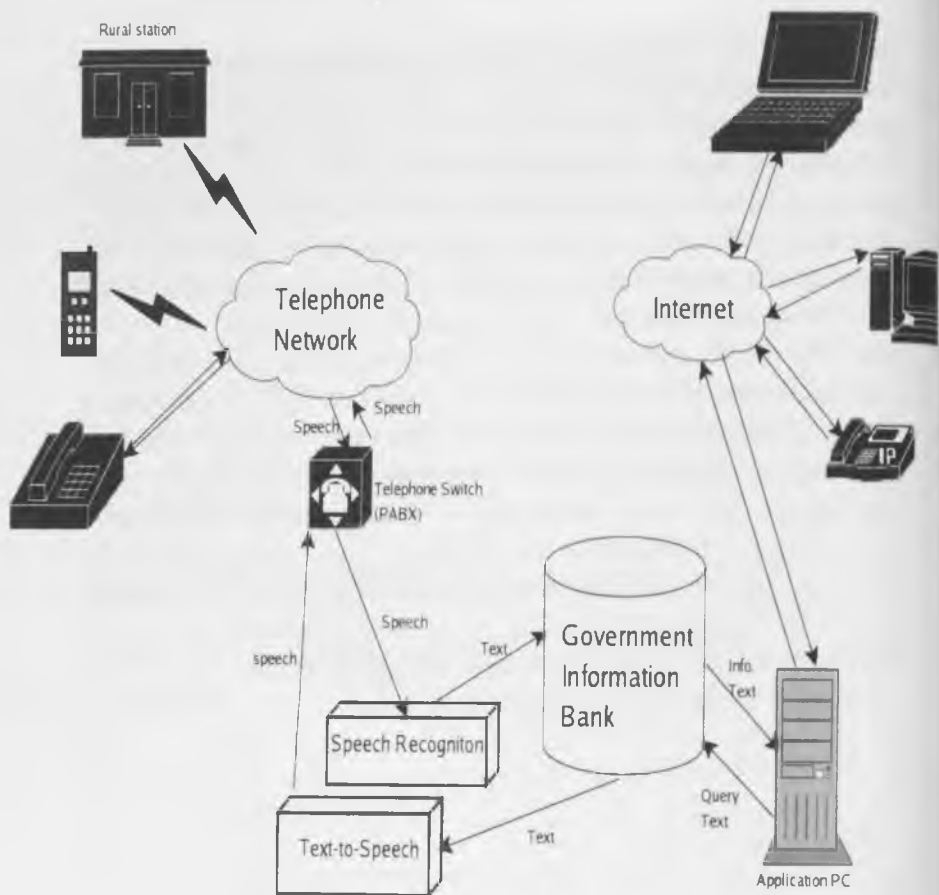
Figure 4: An example of the HLT-ICT integration model

1.1   Description of HLT-ICT model shown in Figure 4

The model shown in Figure 4 illustrates how speech recognition and TTS is connected with the ICT infrastructure. The mobile phone or a telephone handset transmits voice information to the *speechified* government information service. The voice data is then switched to the correct application. The application through speech recognition converts it to text that can use query software to request information from the *Government Information Bank*. The query software is the technique of enveloping

information from an application computer or device in such a manner that the *Information Bank (database)* can understand what the user is looking for. Once the information is found, the TTS engine then coverts the found information into speech, which is transmitted back to the user through the telephony framework (Flood 1997). The same procedure is followed if a caller does not own a landline handset or mobile phone in which case the *rural station* can be used (see figure 4).

Urban people who have access to the Internet can also access the same information through the Internet by making use of the application PC (webserver) (see figure 4). When universal resource locator (URL) address is typed on the Internet browser, the information traverses through the Internet to the webserver, which then sends the request to the *Government Information Bank*. The answer to the request is returned in the form of text and sent back to the users through the Internet. Both the Internet users and mobile/telephone users are able to access information in this case. The non-internet users rely on their language and listening skills while the Internet ones have to know how to browse information from the Internet.

### 3.2 *The design process of speech-based application*

The preceding sections of the paper presented suggestions and theoretical justification on how HLT can empower illiterate members of society through the use of ICT's. Supposing that the government would like to implement automated speech recognition and a TTS- based service in Maseru where people can call and automatically access information in their own language, those who have mobile phones, landline telephones and Internet would have no problem at all. The remote parts of Lesotho can have call stations (rural stations in figure 4) at local government offices or clinics.

The design of this model requires an installation of a computer that *understands* spoken Sesotho. For this

computer to hear speech in Sesotho, the speech recognition system that recognises Sesotho is required. This project therefore requires the design and implementation of such a system. Section 5.2 briefly explains this process and what has been achieved so far.

Experimental setup for the proposed visionary model
Using the model proposed in section 3 above, the experimental setup in figure 5 was designed and implemented.
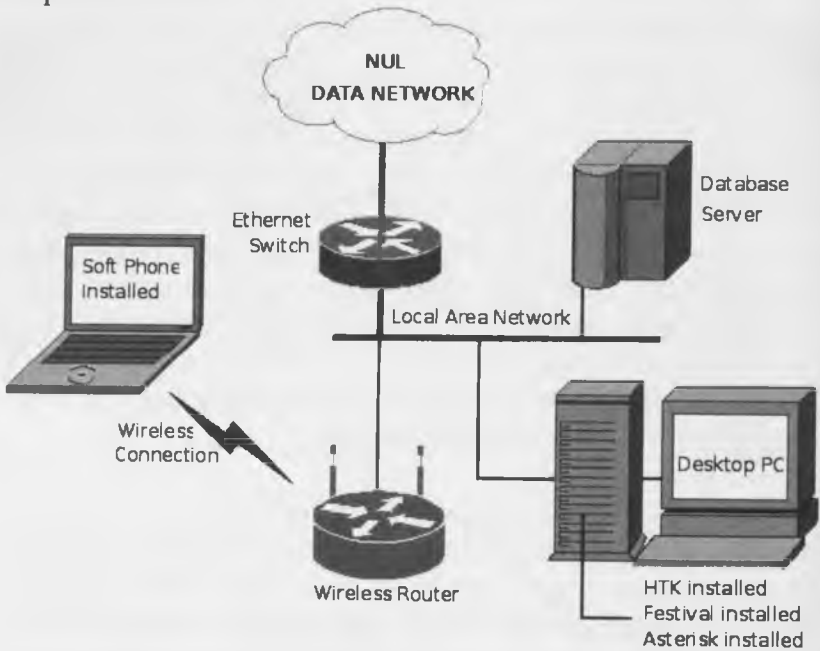
Figure 5: Laboratory setup for the first set of experiments for the visionary HLT-ICT model

Since the paper focuses on how HLT interacts with ICT the authors saw it appropriate to first implement the HLT components of the model. It should be noted that figure 5 does not fully simulate the visionary model described in section 3. It only illustrates the HLT related components.

The components of the experimental setup illustrated in figure 5 are as follows:

4.1 *The speech recognition module*

HTK (Hidden Markov Model Toolkit) (http://www.htk.eng.cam.ac.uk) is the speech recognition software that contains the English Language models, whole-word English dictionary and also the phonetic dictionaries. This software has been installed on the "Desktop PC" as illustrated in figure 5. This "Desktop PC" is a Linux machine with normal office computer specifications. These specifications may require improvements at a later stage. The English models, however exhibit very low recognition performance for Lesotho English accent and therefore new speech acquisition was done where recordings were carried out using a desktop microphone. The new speech files were then included for improving the existing HTK models. Furthermore, a few Sesotho words were added in the HTK dictionary and tested. The ultimate goal is to covert the whole HTK dictionary into a Sesotho dictionary.

4.2 *The TTS module*

Festival is also installed on the "Desktop PC" to represent the TTS module in figure 4. *Festival* is installed so that it can convert text from the "Database server" (*Government information bank*) to speech, which is transmitted to the caller through telephony infrastructure. This module is necessary for demonstrating how HLT integrates with ICT in figure 5. However, not much work in terms of preparing *Festival* to read Sesotho text has been covered. The *Festival*-based Sesotho TTS will eventually be implemented as part of the model in figure 4.

4.3 *The ICT and telephony module*

The telephony infrastructure is simulated by the software called *Asterisk* (http://www.asterisk.org). *Asterisk* software is also installed on the "Desktop PC". Its integration with

both the TTS and HTK is done through its configuration files. The ICT infrastructure already exists at the National University of Lesotho (NUL) campus. The laboratory where experiments are carried out is connected to the NUL data network through the Ethernet switch and cables. The wireless router is connected to the switch forming local area network (LAN) with both the "Desktop PC" and "Database server". The notebook (laptop) which simulates a cellular phone connects to the LAN wirelessly. The LAN represents the ICT infrastructure. The softphone called *X-Lite* (http://www.counterpath.com/x-lite.html) has been installed on the small workstation. This phone uses the normal LAN for communicating with the "Desktop PC" which fetches information from the "Database server" through the same network.

4.4   *Integration of the modules*
The HTK has been configured and tested separately because the *Asterisk* software does not support it. For this reason the HTK-Asterisk interface script (e.g. C++, Perl or Bash program) will be developed as the project progresses. Speech recognition is a very large and broad field of research and it will take the authors sometime to fully accomplish the state-of-the-art speech recognition that will sufficiently illustrate the HLT-ICT integration. For now the authors are focussing on perfecting the HTK based speech recognition engine.

Festival is integrated into Asterisk using the conventional configuration files that come with the *Asterisk* software. The *X-Lite* softphone has also been chosen because it is supported by the *Asterisk*. With all these pieces of software and hardware integrated, the baseline system for our model has been set.

5. Accomplished tests and experiments
This section describes the undertaken laboratory experiments and tests, which form baseline setup essential

for the larger HLT-ICT model. First, the description of the functions and tests of the proposed experimental setup in figure 5 is given and secondly the speech recognition experiments and tests are presented.

5.1 *Functionality of the experimental setup*
In order to test the functionality of the experimental setup in figure 5, we designed its platform architecture in figure 6.
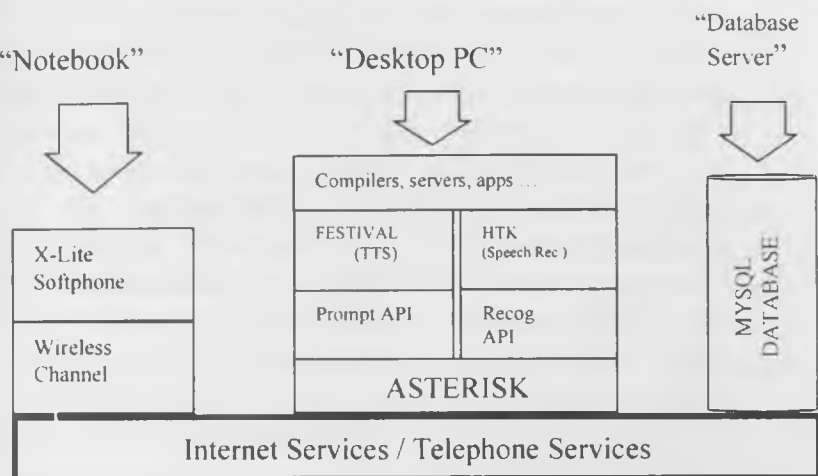

Figure 6: HLT-ICT model platform architecture

Figure 6 shows how each physical computer or network component interconnects with other parts of the HLT or ICT model illustrated in figure 5. In figure 6, the *prompt API* (application interface) is the program that connects *Festival* with *Asterisk* and the users (callers). Similarly the *Recog. API* (Speech Recognition interface) is used by the system to acquire what a caller says to the system, and sends it to the HTK-based speech recognition. The *Recog. API* is not yet implemented. Instead, we have used dual tone multi-frequency (DTMF) (http://www.voip-info.org/wiki/view/Asterisk+DTMF) medium to replace speech recognition engine. The DTMF is usually found on the menu list of a typical interactive voice response (IVR)

system. For example if one calls an automated mobile phone company call centre, a typical response includes: *"press 1 to recharge, press 2 for help ... "*.

The following steps describe how the system is tested:

1. Through the headphones the caller uses the *X-Lite* phone to call the number (e.g. 2000) that is set in the *Asterisk* configuration file called *extensions.conf* .
2. *Asterisk* uses one of its default welcome audio messages to prompt the caller with the menu (e.g. *Press 1 for information and 2 to exit*). The caller makes a selection on the *X-Lite* phone keypad. The selection is transmitted to the *Asterisk* as a DTMF signal.
3. Depending on the caller's selection, the *Asterisk* uses the *extensions.conf* file to fetch text information from the MySQL database via the Ethernet connection. The same file then invokes the prompt API to provide the audio version of the text. This audio is played back through the LAN back to the headphones of the caller and thus completing the call.

5.2   *The speech recognition experiments*

As already mentioned in section 4, the HTK-based speech recognition system has been developed but not yet integrated into the platform architecture in figure 6. The system has similar components as those in figure 1. The aim of the experiment was to ensure that the baseline system works. We undertook the following steps:

1. We recorded about 200 utterances from two people with the aim of achieving high recognition accuracy in accordance with speaker-dependent speech recognition system theory (http://htk.eng.cam.ac.uk; Rabiner and Juang, 1993). Theoretically, 200 utterances from as many people as possible are required to create the initial language model for

speech recognition (http://htk.eng.cam.ac.uk; Rabiner, 1989).

2. We created a lexicon with single English digits as well as four digits that resemble our University's office telephone extensions. This is the simplest lexicon that is recommended in the HTK manual. For a large vocabulary speech recognition, the linguist must ensure that utterances encompass the whole spectrum of sounds of a language. This will be done at a later stage.

3. With the available linguistic information, we created the grammar that allowed each of ten digits to be used individually or in a group of four.

4. We incorporated the recorded acoustic information into the HTK acoustic model for pattern matching.

5. We used the microphone to test the system with the words that appear in the speech grammar and over 80% of the digits and numbers were correctly recognised.

Table 3 summarises progress on the necessary software installations.

Table 3: Summary of laboratory work in progress

Software	Purpose	Progress
Asterisk and Local Area Network	Telephony and ICT infrastructure	Installed and Fully Functional
HTK	The Speech Recognition System	Installed and still at the testing stage
Festival	The TTS System	English text is fully converted and the Sesotho linguistic information is not yet added to the system
IP Soft phones (X-Lite) (IP =Internet Protocol)	Simulates cell phone and landline handset	The IP phones are being installed and are working

Conclusions and Future work

This paper has described the visionary HLT-ICT model which has a potential to benefit underprivileged African countries including Lesotho in various ways. For example the public can use this model for accessing government information without having to travel long distances and waiting endlessly in the queues. The paper has also illustrated how Linguistics is applied in the implementation of speech technology in the form of speech recognition and TTS.

The paper has also provided the design of the simulation environment that can be used to run tests in the building of the real HLT-ICT model. Presently the authors have tested the basic platform and technologies, which are available as free and open source software. The outstanding software and hardware components necessary for the real HLT-ICT model will be added as the experiments progress. Eventually the HLT-ICT infrastructure and relevant HLT-

References

Black, A.W., Taylor, P. and Caley, R. (1999) "The Festival Speech Synthesis System". *System documentation Edition* 1.4, University of Edinburgh.

Cloete, L., Barnard, E. and Patel, H. (2004) "Initial Experiments on the Effectiveness of Telephone Access to Government Services". *Lecture notes in Computer Science.* Berlin:Springer.

Dutoit, T. (1997) "High-quality text-to-speech synthesis: an overview". *Journal of Electrical & Electronics Engineering: Special Issue on Speech Recognition and Synthesis* 17 (1), pp.25-37.

Flood, J.E. (1997) "Telecommunication Networks". IEE Telecommunications Series 36.

Louw, J.A., Davel, M. and Barnard, E. (2005) "A General Purpose IsiZulu Speech Synthesizer". *South African Journal of African Languages,* 25 (2) pp. 92-100.

Naidoo, S., Barnard, E., and M. Davel (2005) "Telephone-based speech interfaces in the Developing World, from the Perspective of Human-human Communication". The 14th International Conference of the African Language Association of Southern Africa (ALASA), Johannesburg.

Owens, F.J. (1993) *Signal Processing of Speech.* London: Macmillan.

Rabiner, L.R. (1989) "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, 77 (2).

Rabiner, L. and Juang, B. (1993) *Fundamentals of Speech Recognition.* PTR Prentice Hall

Riley, T.B. (2005) "e-Government and the Digital Divide: A Commonwealth Perspective". *Networking the Commonwealth for Development,* NDD New Media Ltd. pp. 53-57.

Richards, J.C., Platt, J. and Platt, H. (1992) *Longman Dictionary of Language Teaching and Applied Linguistics,* 2nd Edition. Essex: Clays Limited.

Spio-Garbrah, E. (2005) "Rural Communications: The Key to Bridging the Digital Divide". *Networking the Commonwealth for Development*, NDD New Media Ltd. pp. 21-24.

Tucker, R. and Shalonova, K. (2004) "The Local Language Speech Technology Initiative – Localization of TTS for Voice Access to Information, Crossing the Digital Divide Shaping Technologies to Meet Human Needs". SCALLA Conference, Nepal.

T-Systems, (2004) "Cloning the Perfect DNA for Agent Julie: Speech Recognition Technology Comes of Age". *Convergence Magazine: Special issue on Age of Innovation and Sustainability Supplement*, 5 (4), pp.8-16.

Asterisk DTMF at voip-info dot org, http://www.voip-info.org/wiki/view/Asterisk+DTMF

Human Language Technologies (HLT) – Projects, http://www.meraka.org.za/hlt_projects.htm

International Phonetic Association, http://www2.arts.gla.ac.uk/IPA/ipa.html

Lesotho Telecommunications Authority, http://www.lta.org.ls

Open Source Speech Recognition Software from University of Cambridge, http://htk.eng.cam.ac.uk

Open Source Soft PABX and IVR System, http://www.asterisk.org

Open Speech Recognizer, http://www.nuance.com/products

Poverty Reduction Strategy, http://www.lesotho.gov.ls/articles/2005/

State Information Technology Agency, http://www.sita.co.za

X-Lite softphone (http://www.counterpath.com/x-lite.html