

# The Local Language Speech Technology Initiative

## – localisation of TTS for voice access to information

Roger Tucker, Ksenia Shalnova

*Outside Echo, UK*

[roger@outsideecho.com](mailto:roger@outsideecho.com), [ksenia@outsideecho.com](mailto:ksenia@outsideecho.com)

### **Abstract**

*Phone-based services in local languages would allow widespread access to medical, market pricing, weather and other timely information, as well as email and on-line transactions. Such services are dependent on two types of language technology – text to speech (TTS) to deliver information, and automatic speech recognition (ASR) to access it and control its delivery. Both are complex technologies that require both linguistic and specific technical expertise to develop, and are mostly only available for the world's major languages, and at considerable cost. The Local Language Speech Technology Initiative (LLSTI) is solving this problem by bringing together motivated groups in developing countries, providing tools, expertise, support and training to enable first TTS and then ASR to be developed in their own local languages. The systems developed are made freely available under an open source license, with the aim of building a community of interest to further improve the systems and minimize the barrier to their actual deployment. LLSTI also aims to work in partnership with organisations interested in deploying phone-based services.*

### **1. Introduction**

It is increasingly being recognised by governments, technology companies and development agencies that Information & Communications Technology (ICT) will play a crucial role in growing the economies of developing countries. The problems peculiar to these countries are:

- language - most people do not speak English. Their first language, although spoken by a very large number of people, is typically poorly supported by ICT.
- literacy - in the world's 50 poorest nations, almost half have under 50% literacy. Even some of the world's richest nations (Singapore, UAE) have serious levels of adult illiteracy (15,25%).
- internet access - either not available or prohibitively expensive. Exacerbated by a general lack of IT training.

Information services in local languages with a voice-enabled interface have the potential to solve all three problems. The access device can be low-cost and simple to use – typically a phone, but could be a low-cost PDA such as the Simputer (<http://www.simputer.org>), or even a PC in a community telecenter. However, this requires speech technology to be available in the local languages of the world. Two types of language technology are needed – text to speech (TTS) to deliver information, and automatic speech recognition (ASR) to access it and control its delivery. Of these, TTS is the more urgent technology needed because:

- Voice services can manage without ASR through the use of multiple access numbers for different services, DTMF keys or touch-screens in the case of PDAs.
- A single TTS system can cover quite a large region (using a neutral dialect, as is used for broadcasting), but ASR requires extensive data collection throughout the region it is intended for.
- TTS has many other uses –

- *PC/multimedia*: screen readers, language & literacy learning;
- *Industrial*: speaking alarm systems, announcement systems;
- *Assistive*: reading aids for the blind, communication aids for vocally impaired people.

Localisation of Speech Technology is a massive task. For command-and-control ASR, it is mainly the very time-consuming task of collecting and annotating a lot of data. But a good quality TTS system requires deep knowledge of the morphology, syntax and phonology of the language as well as the technical skill to build the TTS system. Because of this, it is the process of building TTS systems in new languages that is the initial focus of LLSTI. Many of the technologies required in a TTS system are also useful for other speech and language-related applications such as Dialogue Systems, Machine Translation and Information Retrieval.

There are a number of commercial TTS companies, all of which are steadily expanding the number of languages they offer according to likely markets. But where the market is unproven and economically poor, there is little hope of a commercial organisation taking the risk. The alternative approach is for local linguists and engineers to use openly-available tools such as Emu (<http://emu.sourceforge.net>) or Pratt (<http://www.fon.hum.uva.nl/praat/>) and open-source code such as Festival (<http://www.cstr.ed.ac.uk/projects/festival>) and Festvox (<http://festvox.org/>) to build up systems in their own languages. By making results available to other people, as in the MBROLA project (<http://tcts.fpms.ac.be/synthesis/>), a community of interest can be formed, with different people working on different parts of the system, according to their own expertise and interest. LLSTI is committed to enabling and supporting this approach.

## 2. Approach

Through a range of consultations over the last two years or so, a considerable number of institutions were identified that were either already working on aspects of local language text-to-speech for developing countries, or were considering becoming involved. These organisations had been working in relative isolation, but it became clear that there was a genuine willingness to combine their respective efforts into a more co-ordinated, open source approach that is both scaleable and affordable in developing countries. LLSTI is currently working hard to pull these various strands together in a co-ordinated way, and develop the core open source tools to the point at which non specialist local institutions can pick them up for their own local languages.

We have surveyed the languages and scripts used worldwide and identified the problems which will be encountered in building TTS for them, putting what information is available in a database. Our paper at the Budapest workshop [1] summarised the findings of this work in relation to the South Asian languages. This database has enabled us to scope out the tools and approaches needed to produce TTS in *all* the world's written languages.

The essence of the LLSTI approach is as follows:

1. provision of a toolset and training program for linguists & software engineers who are interested in creating a TTS system in a local language.
2. recruiting participants for the scheme from developing countries
3. arranging training of participants in the tools and techniques needed
4. donating equipment to do the data collection and transcription work (*through eg HP philanthropy*)

5. support for participants whilst they do the work
6. compiling the results into a publicly-available central database & open-source code repository.

The toolset is an extended version of the Festival and Festvox tools.

## 2.1 Tools

We are extending Festvox and Festival by:

1. Adding a Morphological Decomposition (MD) framework to Festival. This will also support rule-based chunking & phrasing. This work is being done at IIT Hyderabad.
2. Defining evaluation procedures to help non-experts assess & refine what they have produced.
3. Improving the quality of the signal-processing, first by improving the existing modules, and then by adding a Harmonic plus Noise Modelling (HNM) module to improve the quality for larger pitch modifications.
4. Adding a full-featured Unit Selection System (work at CSTR Edinburgh).

## 2.2 Local Languages

In addition to the tools, LLSTI is also open-sourcing the various languages it produces. At present four languages are being produced by the LLSTI partners:

Hindi (HP Labs India)

Tamil (IISc Bangalore)

isiZulu (CSIR South Africa)

Ibibio (University of Uyo Nigeria & University of Bielefeld, Germany)

## 2.3 Partners

From the beginning, LLSTI has aimed for global coverage. The initial stage is concentrating on forming strong partnerships in two countries in two different continents: India and South Africa. The idea is that the partners in each area will work closely together, with a slightly looser connection between the areas. Each partner organisation is in a different position regarding expertise, but each is strongly committed both to local language speech technology and also to the open source approach.

Not all the project workers have a relevant background – for instance the graduate student from Nigeria has absolutely no experience of speech technology at all. Ultimately the fundamental value proposition of LLSTI is that a motivated engineer/linguist combination with no prior background can produce a usable local language TTS system in one year.

## 2.4 Training and Support

We are currently developing a methodology for training and supporting local people who don't have the complete skill set needed. The partners range in their skills – one has no prior experience of speech technology at all; some have only a little experience of TTS; ourselves (Outside Echo), IIT Hyderabad and HP Labs India have significant experience in TTS.

So far we have run two training courses; a formal two week course for 15 students with an additional aim of finding the best two students to appoint as project workers, and an informal one week course focused on the needs of the two particular project workers already appointed. Both formats were very successful.

As LLSTI scales up to include more partners, the way training is done will have to change. One direction is to make it more web-based. Another is to decentralise the training, so each region could have its own training centre eg South Africa for the African Continent, India for South Asia etc., run by the local partners. As the initiative grows, the number of partners able to train will also grow, enabling the whole program to scale.

## **2.5 Sponsorship for Specific Languages**

It is likely that as more people hear about the initiative, requests for TTS in specific languages will arise. Already, Oneworld International have specifically asked for Tamil & Swahili, with other languages likely in the near future if successful applications are built from these first two. This provides the opportunity for specific funding of those languages, to suit the particular sponsor's requirements. Sponsorship of languages is one way that LLSTI can grow – and sponsorship can come from a wide range of sources, from agencies with a desire to reach into a particular local language community through to commercial operators who can see a new business opportunity.

The success of LLSTI for any particular language will depend on the enthusiasm, focus and commitment of the participants, and the creation of an interest group in that language community that attracts more participation later on – maybe with government or other funding. One problem of targeting specific languages is to find an individual or institution that has a genuine interest, rather than just a financial motive.

## **2.6. Long Term Sustainability**

At present LLSTI is funded by the Department for International Development (DfID) in the UK and the International Development Research Centre (IDRC) in Canada. The ability for LLSTI to sustain itself in the medium to long term is very important. There are various models being considered, but we are definite that joining LLSTI should cost nothing, therefore the subscription model employed by (eg) the LDC is not possible. Thus the central LLSTI management will probably need to be supported either through continued donor support or through making a consultancy business from enabling LLSTI partners to exploit their work commercially.

However, the management costs will remain level as LLSTI scales up to more and more languages, through greater involvement of partners and through a streamlining of the operational methods. Although at present the local languages are being paid for partly by the LLSTI funders, we are aiming to broaden the sources of sponsorship for the local languages themselves, which could come from global NGOs (UN, WHO etc.); local governments; local research funding; mobile operators; (mobile) service providers.

## **2.7 Publicity & Dissemination Strategy**

Key to LLSTI's success is good publicity and successful deployment of the technology. It is being publicised both throughout the Human Language Technologies communities and organisations involved in ITC-based development projects.

Local publicity and dissemination is also carried out by each individual partner, each of whom is investing a certain amount of their own resources and IP into the initiative, and has a vested interest in the success of the initiative. It is in everyone's interest that LLSTI attracts as much attention as possible, both to encourage further participation and also to find potential users of the technology.

### 3. LLSTI TTS Framework

Figure 1 shows the way we are putting together the TTS systems for all our languages. For those readers interested in a bit more detail, each of the modules is explained in the appendix along with our strategy for constructing that module. The modules in italics are not part of the present developments, but would ideally be part of any good TTS system.

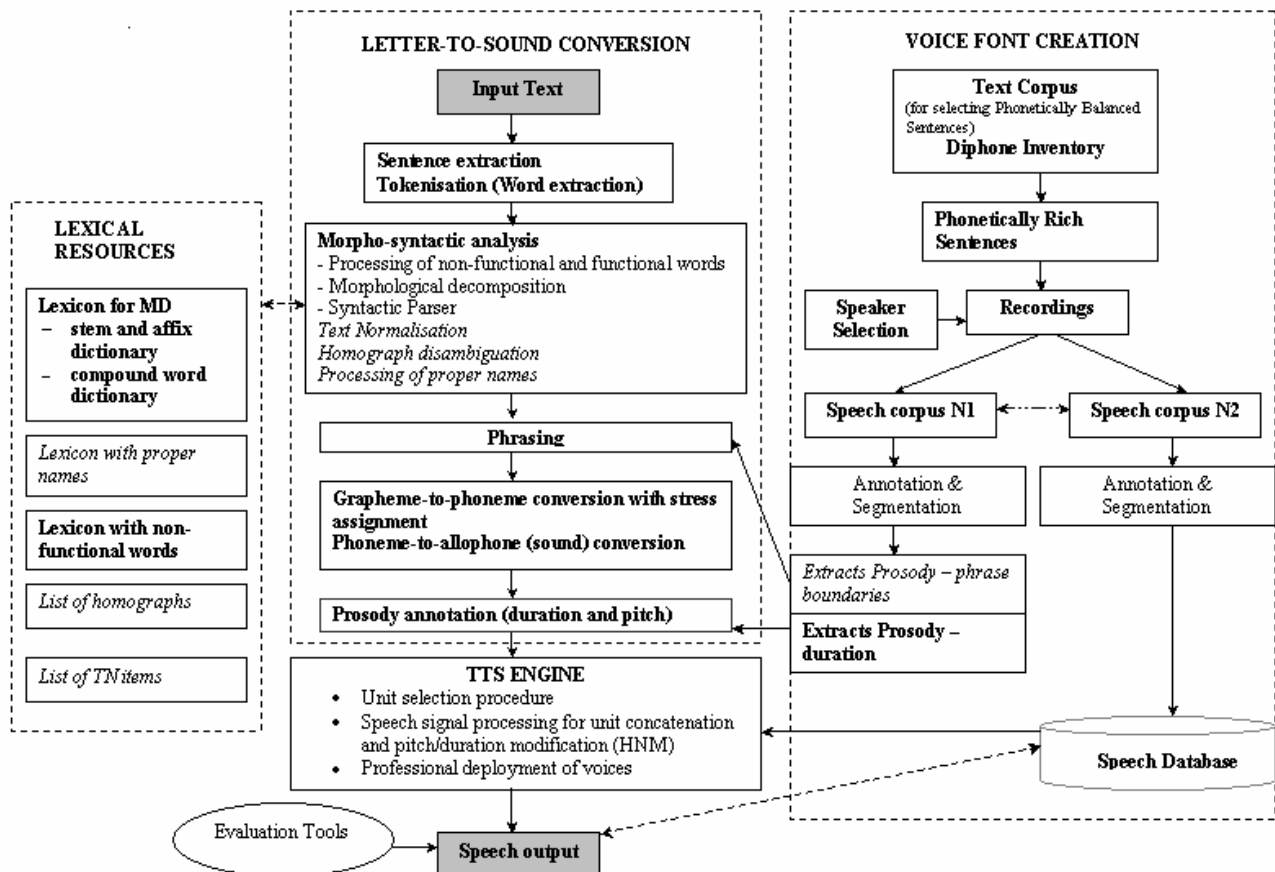


Figure 1. TTS Framework

### 4. Evaluation

Deployment of the TTS produced under LLSTI is key to the vision. HP Labs India (a LLSTI partner) are already piloting a voice service for handling ticket availability enquiries for the Indian railways, which uses pre-LLSTI limited-domain Hindi TTS. Oneworld are planning to pilot Tamil weather and market pricing information. The take-up statistics of these and other expected deployments will give a general indication of the value of TTS-based voice services.

However, we are keen to evaluate the TTS-based voice interface on the less-educated, less-literate people we are trying to reach, and the best-placed partner to do this is CSIR in South Africa, who have a number of existing projects where the isiZulu TTS can be introduced and tested in specific user-trials. These are:

- Digital Doorway <http://www.digitaldoorway.co.za/> - computer usage is monitored (video cameras activating whenever users utilise the technology) providing a very good test environment and allowing analysis of user reaction to the TTS system in detail.

- Manguzi (<http://www.cda.co.za/Manguzi/index.html>) - an established telecentre in a remote rural community in KwaZulu-Natal which has been able to achieve community buy-in and is well-utilised. This project provides access to a large group of Zulu speakers from a rural background.
- E-gov service delivery - an experimental project aiming to obtain a better understanding of issues of culture, illiteracy and the effect of limited exposure to technology when developing telephone-based information systems.
- Multi-lingual weather information system – a new project utilising machine translation.

The focus is likely to be on the E-gov service delivery, though this has yet to be firmed up.

Evaluation of the unrestricted-domain TTS system on target users is very important. We have yet to discover how tolerant technology-shy or illiterate people are to both the mistakes made by TTS and the unnaturalness of the voice, which will determine the quality required for an effective deployment.

## 6. Conclusion

LLSTI is still in its early stages. Its strength lies in the commitment of each of its partners to the vision of good-quality, open-source, freely available TTS. The highly competitive model for TTS in the main Western languages is exchanged for a genuinely open and co-operative model, more likely to produce the quality required. LLSTI focuses not only on the technology itself, but also how it can be deployed and what is needed to make those deployments successful.

## Reference

[1] ‘South Asian Languages in Multilingual TTS-related Database’, *K. Shalnova, R. Tucker*, EACL Workshop on Computational Linguistics for the Languages of South Asia - Expanding Synergies with Europe, Budapest, April 2003, pp 57-63

## Appendix – Description of TTS Modules

Modules	Purpose of a Module	Development strategy; required linguistic resources or skills
<b>LETTER-TO-SOUND CONVERSION</b>	Converts graphemes into allophones (or phones) regarding their prosody characteristics; each allophone (or phone) should have a corresponding speech unit in the Speech Database.	The process is normally iterative – at least 2 versions are required.
<b>Sentence extraction</b>	Extracts sentences from a text.	based on punctuation marks (.?!...)
<b>Tokenisation</b>	Extract words from a sentence.	based on spaces between words
<b>Morpho-syntactic analysis</b>	In TTS systems Morpho-syntactic analysis can be required for generating proper G2P rules, stress and tone assignment, homograph disambiguation and intonation modelling.	
1. Processing of non-functional and functional words	Generation of phonetic words - combining non-functional words with functional words (e.g., I want to_do this_task)	Lexicon with non-functional words (prepositions, conjunctions and particles) is required (with the indication of the attachment position of a non-functional word: either before or after the content word). As the list of non-functional words is

<b>Modules</b>	<b>Purpose of a Module</b>	<b>Development strategy; required linguistic resources or skills</b>
		highly restricted, it seems feasible to generate it manually.
2. Morphological Decomposition	The input is a wordform extracted from a sentence. The output of MD module is a wordform with the marked morpheme boundaries with their grammatical characteristics.	The following input is required: <ul style="list-style-type: none"> <li>• Stem and affix lexicons including POS (part of speech) information</li> <li>• Morphotactic rules and a paradigm table</li> </ul> Morphological learning tool in the current phase will be tested only for Hindi.
3. Compound word decomposition	Decomposition of a compound word into several single words (is used for proper G2P, stress assignment etc.)	Compound word dictionary OR Syllable-based algorithm using N-grams
4. Syntactic Parser (rule-based)  Shallow parser will probably be developed for this project.	Providing the underlying structures of a sentence.	Provide grammars for each language.
<b>Phrasing</b>	To subdivide a sentence into phrases.	The current Festival tool uses only punctuation marks for this purpose. As punctuation marks do not always mark the phrase boundaries, our tool will be based on shallow parser.
<b>Grapheme-to-phoneme and Phoneme-to-allophone (or phone) conversion.</b>	Providing automatic conversion of graphemes into the sequence of allophones (or phones) with lexical stress and tone assignment.	Grapheme-to-phoneme and Phoneme-to-allophone rules. Morpho-syntactic analysis if necessary (e.g., for schwa deletion in Hindi, for tone assignment in Ibibio etc.)
<i>Text Normalisation</i>	<i>Providing identification and proper pronunciation (G2P rules) for abbreviations, Acronyms, E-mail/URL, Digit strings: currencies, dates, telephone numbers, bank accounts, time etc.</i>	
<i>Homograph disambiguation</i>	<i>Disambiguation of heterophonic homographs</i>	<i>List of homographs and text corpus including these homographs. N-gram or HMM algorithm can be used for training. Rule-based approaches using syntactic information can also be used.</i>
<i>Proper names processing</i>	<i>Identification and generating proper pronunciation for proper names.</i>	<i>Special heuristic techniques as "Gate" from Sheffield University can be used for proper name recognition in a text. Pure statistical approaches can also be used.</i>
<b>Prosody annotation (duration and pitch)</b>	Assignment of proper prosodic parameters to each allophone.	ToBi, Fujisaki, (TBD for each language)
<b>TTS ENGINE</b>		
<b>Unit Selection Procedure</b>	Searches for the most appropriate diphone among a set of candidates that has the best match to the target diphone according to phonetic conditions at the transcription level (Target Cost) and acoustic level (Concatenation Cost).	The algorithm can be tuned either to a particular language or a speaker by assigning different weights both for Target and Concatenation cost. Has certain connection to the phonetically balanced algorithm.

<b>Modules</b>	<b>Purpose of a Module</b>	<b>Development strategy; required linguistic resources or skills</b>
<b>Speech signal processing for unit concatenation and pitch/duration modification</b>	Modifies duration and pitch values according to the defined values in the input transcription sequence and also smoothes the spectrum at the diphone boundary.	The naturalness of the generated speech is the main criteria for selecting an appropriate method.
<b>Professional deployment of voices</b>	Festival is slow, so a faster version needs to be developed for multi-channel real-time deployment.	
<b>VOICE FONT CREATION</b>		
<b>Text Corpus (for phonetically balanced sentences) or Dictionaries</b>	A corpus of transcribed texts (either obtained with G2P or manually) from which phonetically balanced sentences will be chosen.	Random texts.
<b>Diphone Inventory</b>	Creating a sufficient set of diphones required for synthesizing possible speech sequences.	A set of possible allophones (phones) generated by G2P + phonotactic constraints are required for this module. There can be speaker specific G2P rules, i.e. the set of diphones for different speakers can be slightly different. As in developing G2P, diphone creation is also an iterative procedure.
<b>Phonetically Rich Sentences</b>	Selecting a set of sentences (scripts to be pronounced by a speaker) that include a sufficient number of diphones to be used in data-driven TTS systems.	<ol style="list-style-type: none"> <li>1. Corpus of transcribed texts (obtained either with grapheme-to-phone converter or manually).</li> <li>2. Diphone (allophone) inventory</li> <li>3. Frequency syllable or allophone characteristics derived from the corpus of transcribed texts. The frequency characteristics are required for providing the fullest set of possible diphones for each frequent target diphone.</li> <li>4. Phonetic features to be taken into account during the search process (position in a word, lexical stress position, phonetic context etc.).</li> </ol>
<b>Speaker Selection</b>	Selection of a proper speaker.	Requires a set of speaker's characteristics to be evaluated.
<b>Recordings</b>	Recording scripts.	
<b>Annotation &amp; Segmentation</b>	To segment and annotate the recorded speech material into allophones (phones). Annotation should include features to be taken into account during unit selection procedure (e.g., syllabic boundary, lexical stress etc.).	Requires phonetic skills.
<b>Extracts Prosody – duration</b>	Assignment of sound durations.	Automatically builds duration CART-trees on the basis of the annotated speech corpora (annotation should include features that are to be taken into account during training).
<b>EVALUATION TOOLS</b>	Tools for evaluating the quality of the synthesized speech.	Minimal pairs intelligibility test, detecting flawed acoustic units etc.