Text-To-Speech Technologies for Mobile Telephony Services

Paulseph-John Farrugia

Department of Computer Science and AI,
University of Malta

Abstract. *Text-To-Speech* (TTS) systems aim to transform arbitrary[1] textual input into spoken output. At first glance, this may seem a relatively simple task of determining the phonetic sounds of the input and outputting a corresponding sequence of audible signals. However, it is in fact quite a difficult task to produce *intelligible* and *natural* results in the general case. This is due to linguistic and vocalization subtleties at various levels that human speakers take for granted when interpreting written text. In fact, the task requires a considerable grasp of both *Natural Language Processing* and *Digital Signal Processing* techniques.

The potential application of such functionality is varied, including its use for language education, as an aid to handicapped persons, for implementing talking books and toys, for vocal monitoring and other man-machine communication facilities. The area is currently being explored in order to address its application for the Maltese language[2] within the context of the mobile communications industry.

This paper's main aim is to provide a brief overview of current TTS approaches and techniques, and the way these may be implemented. For further insight, reference should be made to the selected bibliography.

1 Introduction

In general, the transduction process from text to speech is carried out through a sequence of readily recognizable steps that provide an increasingly detailed (*narrow*) transcription of the text, from which the corresponding spoken utterance is ultimately derived. These steps can be divided into two blocks—the *Natural Language Processing* (NLP) block and the *Digital Signal Processing* block. This organization is shown in Fig. 1 (Adapted from [3].)

These steps should not be thought of as 'filters,' but as processes that incrementally augment the information derived from the input and place it on a commonly accessible *Multi-Level Data Structure* (MLDS). The ultimate aim is to derive sufficient information on the MLDS so as to be able to derive an intelligible (i.e. readily understandable to a listener) and natural (i.e. comparable to a human speaker as regards intonation and prosody.)

2 The NLP Block

The NLP block consists of those processes that are responsible for analyzing the text in order to extract a sufficiently detailed *narrow phonetic transcription* that can eventually be used for the

[1] As opposed to 'canned', such as is the case with pre-recorded, concatenated IVR messages.

[2] Although attempts are made to unify the field across languages, TTS implementations need to cater for language specific phenomena.

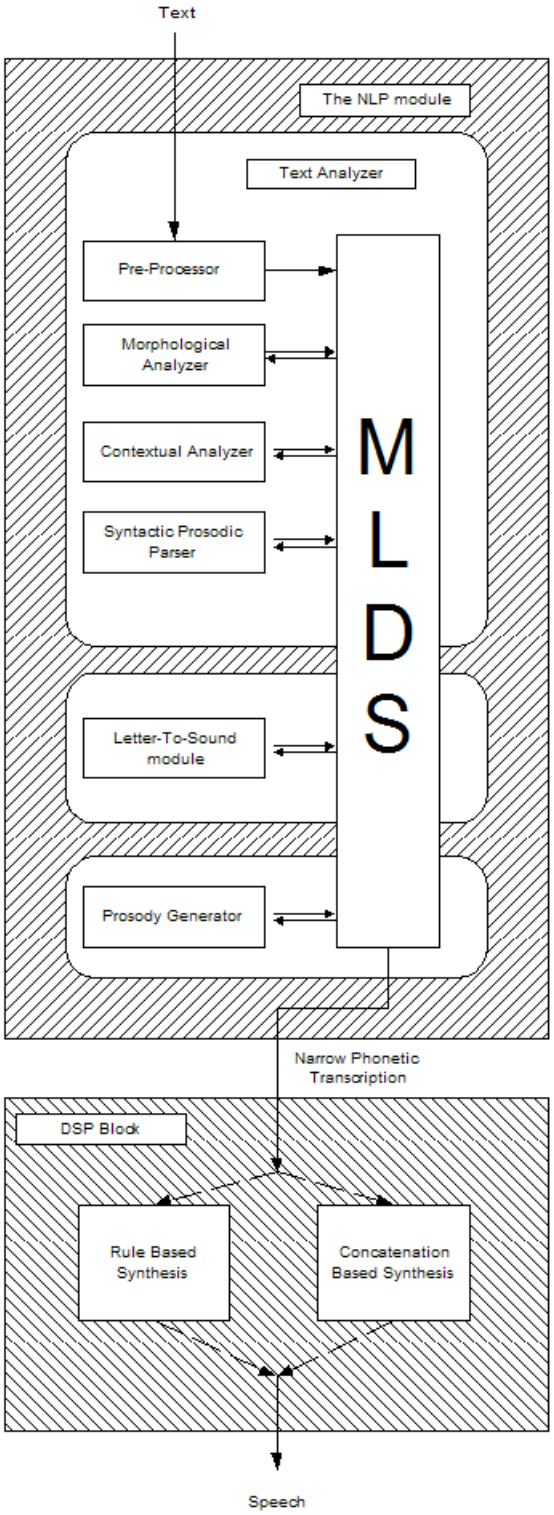

Fig. 1. TTS Processes

DSP, or synthesis, block. This may be considered the more language-specific block, as it needs to derive a concrete description of how the textual input should sound (without actually deriving the corresponding sound per se.)

The processes will now be considered in order.

2.1 Pre-Processor

The pre-processing module is responsible for transforming the text into processable input, that is, into a list of words. Among other things, it is responsible for:

- recognizing and spelling out acronyms, which is not so straightforward when one considers different interpretations of acronyms such as, for example: *CNN*, *SARS*, *CSAW*
- spelling out numbers, taking context into consideration, as in, for example: *25*, *Lm 25.00*, *15 ktieb*.
- resolving punctuation ambiguity as, for example between a full stop identifying the end of a sentence and one at the end of an abbreviation.

2.2 Morphological Analyzer

The morphological analyzer utilizes lexical information in order to derive a morphological parse for each word, and consequently identifies its possible parts of speech.

2.3 Contextual Analyzer

The contextual analyzer considers the surrounding context of the words in order to limit the possible parts of speech to a minimal, likely set.

2.4 Syntactic Prosodic Parser

The syntactic prosodic parser organizes the lexically analyzed word list into identifiable boundaries (sentences, clauses and phrases.) This text structure will then be utilized in order to derive an appropriate prosodic representation.

2.5 Letter to Sound Module

The *Letter to Sound* (LTS) module, (also referred to as the *grapheme to phoneme* module) derives a phonetic transcription for the input. Once again, upon first inspection this may seem like a trivial process of looking up entries in a pronunciation dictionary. However, this is in fact not feasible, and the process is a complex one for the following reasons:

- Due to morphology, no pronunciation dictionary could be considered complete in respect of providing an entry for every possible word form. An ordinary dictionary, for instance, usually only provides a pronunciation for a basic word form. Thus, even if a pronunciation dictionary is used, morphophonological rules will be required in order to account for derived words.

- Some words will lend themselves to more than one pronunciation depending upon their part of speech or context. These are referred to as *heterophonic homographs*, examples of which are *sur* (**/sur/** or **/suːr/**) and *bajjad* (**/bɐj-jɐt/** or **/bɐj-ˈjɐːt/**.)
- The pronunciation of a word is affected by its surrounding textual context, in particular at the word boundaries. Hence, it is may not be possible to store an absolute pronunciation for a word independent of context.
- Even the most complete pronunciation dictionary would not contain entries for new words which make their way within a language, or for other classes of words such as proper names.

Two main approaches to this problem can be found. The first, *dictionary-based* approaches, do indeed utilize a pronunciation dictionary in order to provide as much language coverage as possible. Entries are usually stored in the form of morphemes, with pronunciation of derived forms obtained using morphophonemic rules. Unknown words are transcribed by rule. The *rule-based* approaches invert this by transcribing most of the input by general grapheme to phoneme rules, and only keep a relatively small pronunciation dictionary for known exceptions.

2.6 Prosody Generator

The term *prosody* refers to the speech signal properties such as intonation, stress and rhythm. The effects of prosody have a direct influence on how the message represented by the text is conveyed, differentiating, for instance, between a statement or a question (by means of changes in intonation) or providing focus on the subject matter (by means of appropriately placed stress.)

The prosody generator, then, is responsible for identifying the intonational phrases corresponding to the text and assigning the appropriate prosody contour.

3 The DSP Block

The DSP block, depicted in an over-simplified manner in Fig. 1, is, at an intuitive level, the programmatic counterpart of the human speech reproduction organs. It utilizes the narrow phonetic transcription derived from the previous block in order to generate the actual speech waveform that can be audibly reproduced.

Two main approaches are encountered for this task. The first is the *rule based* approach, which attempts to provides synthesis through the modelling of the human speech reproduction process as the dynamic evolution of a set of parameters. This approach has a number of benefits. For instance, it is *speaker independent*, that is, the voice is completely synthetically generated, and can hence be altered at will by modifying the appropriate parameters. However, this approach is complex and takes a long development effort due to the difficulty in deriving the appropriate set of rules.

The second approach is the *concatenative* one, which utilizes a speech database consisting of pre-recorded elementary speech elements (typically *diphones*, two sequential phoneme units) as the basis for synthesis. In practice, the output is generated by concatenating a sequence of elementary speech elements corresponding to the phoneme sequence identified by the NLP block. DSP algorithms are then applied in order to smooth the resulting waveform and apply the intended prosody. By the contrast to the rule-based approach, concatenative synthesis is strictly speaker dependent, as the speech database is generated from recordings of appropriately chosen text by a professional speaker.

4 Research and Application

Research is being carried out in order to apply TTS techniques in order to be able to develop a system for the Maltese language, with a practical application within the context of mobile telephony in providing extended services such as SMS to voice mail. Particular emphasis is intended on the prosody generation phase. The aim is to utilize previous work on TTS for Maltese ([7]) and Maltese prosody ([10]) in order to develop a framework for more natural synthesis results.

References

1. Jonathan Allen. *Overview of Text-to-Speech Systems*, chapter 23, pages 741–790. In Furui and Sondhi [4], 1991.
2. Thierry Dutoit. High-quality text-to-speech synthesis: An overview. *Journal of Electrical and Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis*, 17(1):25–37, 1997.
3. Thierry Dutoit. *An Introduction to Text-To-Speech Synthesis*, volume 3 of *Text, Speech and Language Technology*. Kluwer Academic Publishers, P.O. Box 322, 3300 AH Dordrecht, The Netherlands, 1997.
4. Sadaoki Furui and M. Mohan Sondhi, editors. *Advances in Speech Signal Processing*. Marcel Dekker, Inc., 270 Madison Avenue, New York, New York 10016, 1991.
5. Ivana Kruijf-Korbayová, Stina Ericsson, Kepa J. Rodríguez, and Elena Karagjosova. Producing contextually appropriate intonation in an information-state based dialogue system. *EACL '03*, 2003.
6. Mark Y. Liberman and Kenneth W. Church. *Text Analysis and Word Pronunciation in Text-to-Speech Synthesis*, chapter 24, pages 791–831. In Furui and Sondhi [4], 1991.
7. Paul Micallef. *A Text To Speech System for Maltese*. PhD thesis, University of Surrey, 1997.
8. Hirokazu Sato. *Speech Synthesis for Text-to-Speech Systems*, chapter 25, pages 833–853. In Furui and Sondhi [4], 1991.
9. Richard Sproat, editor. *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachussets 02061, USA, 1998.
10. Alexandra Vella. *Prosodic Structure and Intonation in Maltese and its Influence on Maltese English*. PhD thesis, University of Edinburgh, 1994.