Linguistically Motivated Text to Speech Technologies for Telecommunication Services with Mixed Language Support

M.Sc. Seminar Proposal

Paulseph-John Farrugia

Supervisor: Mike Rosner
University of Malta[*]

December 2003

Abstract

High quality *Text to Speech* (TTS) services are one necessary component for improving man-machine communication interfaces and for enabling *Unified Messaging*, expected to be an important evolution in telecommunications. Briefly stated, the Unified Messaging concept aims to unify different forms of communication media, such as voice, email, fax and instant messaging, providing a single point of access to the user.

To this end, this project explores the use of TTS technologies for the purpose of developing a system capable of converting SMS messages to spoken output, thus opening the medium to other formats of delivery. In the local context, SMS messages are generally written in Maltese, English or, as is quite typical, in a mixture of both languages, and also contain various word contractions and shorthand notation, typical of the SMS lingo.

The aim is to focus on two main aspects of the problems arising. First, the essentially multilingual nature of the input domain, as well as other idiosyncrasies, are taken into consideration, instead of being avoided. The input is not assumed to consist of academically correct Maltese, and a statistical technique is presented for handling the *code switching* of various words.

Secondly, in order to produce more natural speech, a framework for describing and assigning *intonation* and other *prosodic* features of the input language is required. An outline is given of how such a formalism, adequate for the purposes of TTS, can be developed on the basis of previous phonological studies of the Maltese language.

After a brief introduction to TTS and relevant literature, each of the two issues above is explored in greater detail. An overview of relevant resources that can aid the implementation process is then given, followed by an illustration of how the technology can be put to practical use. Finally, the project's main objectives are presented, including a list of expected deliverables.

1 Introduction

TTS systems aim to transform arbitrary textual input into spoken output by way of the automatic phonetization of the sentences to utter ([20]). The fact that the input is assumed to be arbitrary, as opposed to 'canned', such as is the case with pre-recorded, concatenated IVR messages, gives a variety of potential applications for TTS, including its use for language education, as an aid for the visually disabled, for implementing talking books and toys, for vocal monitoring and for other man-machine communication facilities ([19, 20]).

At first glance, TTS may seem to consist of a relatively simple task of determining the phonetic sounds of the input and outputting a corresponding sequence of audible signals. However, this view of TTS as a simple rewrite problem is fallacious, and 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 or spoken

[*] This work is supported by MobIsle Communications Ltd. (http://www.go.com.mt)

Feature	**Example**
Heterophonic Homographs	Phonetizing *bajjad* requires it being distinguished between its use as a noun (*painter*, /**bɐj-ˈjɐːt**/) or a verb (*he painted*, /ˈ**bɐj-jɐt**/).
New Words	Phonetizing new words and acronyms—compare *CNN*, *IEEE* and *CSAW*.
Phonetic Assimilation	The /**b**/ phoneme becomes **[p]** at the end of a word such as *kelb*.
Stress Assignment	Identifying what linguistic elements to stress may require a semantic or pragmatic interpretation of context—compare “Ġorġ *telaq*” vs. “*Ġorġ* telaq.”
Intonation Contour Assignment	Identifying whether to apply an interrogative pitch contour or a statement contour.
Vocal Declination	Modelling the relatively decreasing Fundamental Frequency (F0) of an utterance over long stretches of speech.

Table 1: Vocalisation Subtleties

text. A short and non-exhaustive list of linguistic features out of which implementation difficulties arise is given in Table 1.

The human reading process (an overview of whose intricacies is given in [20]) is quite a complex one. (Figure 1 provides an interesting intuitive example of the facility with which the human mind can easily process even miss-spelled texts.) In addition, modelling the acoustic traits of the human vocal apparatus presents its own difficulties. ([4] provides a comprehensive introduction to acoustic phonetics.) Under closer scrutiny, then, implementing a TTS engine requires considerable use of both NLP and DSP techniques.

The process of generating synthetic speech has been under study for quite some time. A very interesting exposition of its evolution and development can be found in [43]. Starting from a predominantly speech-signal oriented discipline, TTS is now a more structured technology whose study encompasses a substantial amount of linguistic study ([20]), generalization for multiple language support ([59]) and structured mark-up languages ([61, 60, 68]) and APIs ([37, 29]), with resulting systems reaching a high level of output quality. (See, for instance, the online demos at [45], [28], [2], [44] and [55].)

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*) phonetic transcription of the text, from which the corresponding spoken utterance is ultimately derived. These steps can be categorized into two blocks, the *NLP block* and the *DSP block*. This organization is shown in Figure 2 (Adapted from [20], with the DSP block consciously over-simplified as it does not represent a focal point of study within this project.)

The modules within the NLP block should not be thought of as filters, but as processes that incrementally augment the information derived at the current processing stage and store it on a commonly accessible *Multi-Level Data Structure* (MLDS) or set of *Feature Structures* (FSs). From the latter, a corresponding acoustic signal is generated by the DSP block, through *rule-based* (*articulatory* or *formant*) synthesis or *concatenation-based* (*diphone*) synthesis.

While the function of each of these processes can be more or less clearly delineated, the ideal implementation in each case is still subject of active research. Further exposition can be found in [24] and [19], while [54], [27] and [5] provide reports on different candidate implementation techniques for the individual modules and corresponding results.

2 Mixed Language Support

The *Pre-processor* is the first stage of input processing, and is responsible for organizing the input into a standard format that the following modules can process more easily. Amongst other things, it is generally responsible for converting numerals and acronyms into their textual interpretation. In specialized systems, other tasks may be required of it. For example, in an email to speech system, the pre-processor will be required to remove unnecessary email headers and superfluous formatting. (See [9] for an example.)

In this case, the pre-processor will be required to han-

The following text is taken from an email currently being forwarded around the Internet:

> "*Aoccdrnig to rscheearch at an Elingsh uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, olny taht the frist and lsat ltteres are at the rghit pcleas. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by ilstef, but the wrod as a wlohe. Fnnuy how the mnid wroks, eh? . . .*"

Irrespective of the validity of its message, it does provide an amusing example of the facility with which the human mind can easily interpret and read out even misspelled texts.

Figure 1: Reading Misspelled Text

dle issues that arise out of the mixed language nature of the input domain. The main two issues, *code switching* and character encoding support, are discussed below. Another pre-processing task envisaged in this case is the handling of *smileys*, such as `;->`, and other abbreviations typical of SMS lingo. This is expected to be, however, a simple rewrite process from well-known lists of such common shortcuts[1].

2.1 Bilinguality and Code Switching

The Maltese Islands are officially bilingual, with both Maltese and English as official languages of the country. English is generally understood as British English, while an official version of Maltese, often termed as *Standard Maltese* (see [10] and [11]) is commonly taught in schools. The latter is distinguished from dialects of Maltese that may be found in certain villages and localities, which differ subtly at various levels of structure (including lexical, grammatical and intonational.)

In practice, however, this bilingual situation brings forth a heterogeneous mixture of language use. For various geographical and sociological reasons ([10]), the use of Maltese, whether spoken or written, is often interspersed with words or phrases in English[2], a phenomenon known as *code switching*. Conversely, the local use of English is often marked by certain grammatical and phonetic differences, to the extent that some may want to refer to it as a "Maltese English" dialect ([66]).

For an appreciation of the above, consider, for instance, the following (real-life) SMS text message sample:

> *"jaqaw bdejt tibza tixel il car? xorta qajjimt il qattusa u issa qed tajjat wara il bieb. bring that btl of last time plz qalbi :)"*

The extent to which such code switching is acceptable as proper use of the Maltese language (or whether "Maltese English" may be truly considered an English dialect as opposed to a result of a lack of linguistic sophistication) is a matter of near religious debate ([66]). Rather than going into the respective merits of such considerations, a pragmatic approach is preferred whereby this state of affairs is addressed and techniques that can help us solve the issues that arise are identified. In this case, a means is required to identify the switching from one language to another at the word level[3].

In respect of arbitrarily identifying the language of a text, two main trends can be identified, both based on statistical techniques ([31].) The first involves the use of frequencies of occurrence of short words (such as u and ta in Maltese) in order to determine the general language of a given corpus. This is not a directly applicable approach here, since the interest lies in identifying the language of origin of each individual word.

A second technique, which can be used for language ([7]) and subject ([13]) classification, is based on *n-gram* probability matching on characters. In essence, for each of a candidate set of languages, an n-gram

[1] One might consider, however, that smileys are used to convey the emotional states of the sender, and may hence be used as boundary markers for prosodic parsing.

[2] And, to a lesser extent, with words in other languages, most typically in Italian.

[3] Ideally, this should be at the morpheme level, since it is also common to process loan words using Maltese morphological rules. Also, to a certain extent, it may be desirable to phonetize other orthographic elements, such as numerals, in another language, in order to conversely mimic this tendency toward code-switching.

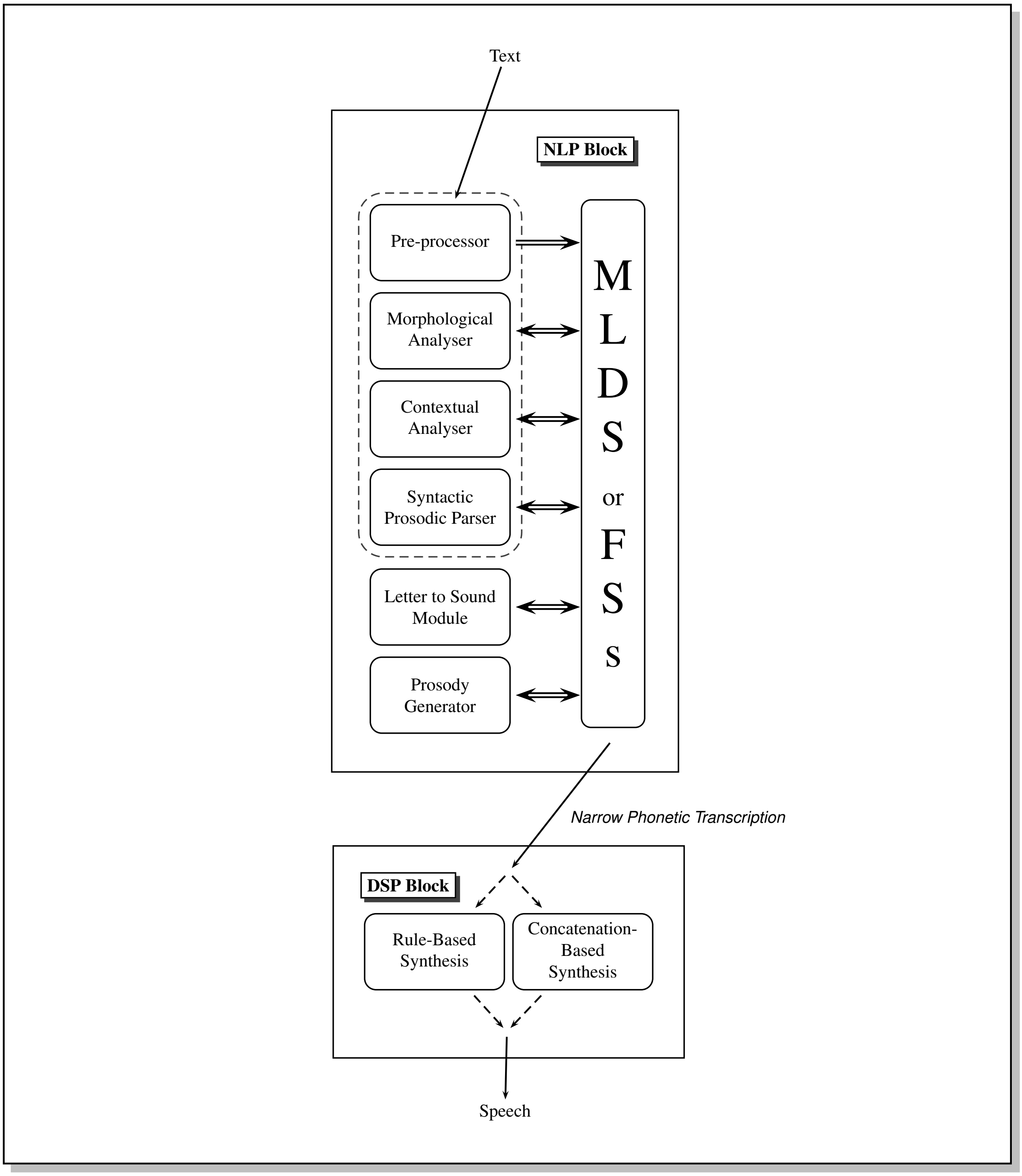

Figure 2: Typical Organization for Text to Speech Processing

probability profile is created from a training corpus. Input text is then classified according to the profile that best matches its n-gram probabilities.

An extension is hereby proposed to handle the issue of code switching. Assuming *bigrams* are being used, the procedure can be expressed as follows. Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages. For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$ be the set containing those characters belonging to the language L_i, taken from a universal set of characters $\mathbf{C}$.

Given any significantly large corpus known to be entirely (or primarily, since the interest here lies in statistical significance) in a language L_i, it is possible to compute $0 \geq P_{L_i}(a, b) \leq 1$, the probability that the bigram ab occurs in free text written in the language L_i, where $a, b \in \mathbf{C}$. The probability of a word $\mathbf{w} = w_1 w_2 \ldots w_k$, where $w_j \in \mathbf{C}$, belonging to L_i can then be defined as:

$$Pw_{L_i}(\mathbf{w}) = \prod_{j=0}^{k} P_{L_i}(w_j, w_{j+1})$$

where w_0 and w_{k+1} are defined as the equivalence class of characters in the set $\mathbf{C} \setminus C_i$, that is, those characters not belonging to language L_i (which include spaces and punctuation.) The most likely classification of $\mathbf{w}$, then, is that it belongs to the language L_i that maximizes $Pw_{L_i}(\mathbf{w})$ over all languages in $\mathbf{L}$. In practice, it would also be possible to take the probabilities of the surrounding words for further disambiguating power.

2.2 Character Set Support

Another issue to be tackled arises from the character set required to encode the Maltese alphabet. Proper writing of Maltese mandates the use of certain characters, namely *ċ*, *Ċ*, *ħ*, *Ħ*, *għ*, *Għ*, *ż*, *Ż*, that are not found in the ASCII character set. Electronic input devices—from keyboards to mobile phones—in the local market are usually intended for an English audience (perhaps supporting some other main European languages), and the support for the full Maltese character set ranges from limited (as is the case with most common computer keyboards in use, for which specific shortcuts may be required) to non-existent (as in the case of some mobile phones.) Given the use of English as a secondary language, this is generally not seen as a major usability issue, and the motivation to have such devices enabled is somewhat limited.

While the advent of Unicode ([64]) is now helping to provide a reference for Maltese text processing (although it still poses some computational issues, as it does not admit the representation of digraphs, such as *ie* and *għ*, as single character units), the situation regarding electronic texts at this point is really split in two. As regards official documents or documents meant for public access, schemes have been utilized to represent the above characters using the ordinary ASCII encoding. The most typical is the use of particular fonts designed for Maltese typesetting (but incompatible with Unicode) whereby the glyphs of the above characters replace the entries for certain punctuation characters. Under other schemes, such as that used in Maltilex ([49]), escape sequences are used instead.

On the other hand, where the presentation of text is not critical, it is often the case that letters containing diacritical marks are written instead with those ASCII character counterparts that do not, which, corresponding to the above, would be *c*, *C*, *h*, *H*, *gh*, *Gh*, *z* and *Z*. Despite what would constitute intrinsic spelling mistakes, a native speaker can immediately identify and interpret the underlying message[4]. However, such text cannot be directly operated upon prior to resolving the resulting lexical ambiguities.

Resolving this issue is an exercise in spell checking, albeit a simplified one. One could envisage resolving it using a threefold approach (starting from the assumption that the word is actually a purely Maltese one.) First, certain re-write rules can be applied. For instance, *c* can be safely re-written as *ċ* in all cases, since the previous character does not exist in the Maltese alphabet. Secondly, use of a dictionary can sort out some entries, as it will list *żarbun* but not *zarbun*. Finally, one could consider using stochastically based re-write rules trained on a corpus written in two fashions, once in the appropriate encoding and one using the ASCII character set only.

3 Describing Prosody

Whilst there are various open research areas within the context of developing a TTS system, a central one is that of properly assigning appropriate prosodic contours. In this section, an overview of this area is given and a need for a prosodic description of Maltese, that is

[4] Essentially, this is a more natural occurrence of the phenomenon illustrated in Figure 1.

also appropriate for computational purposes, is identified.

3.1 Prosody

The term *prosody* refers to speech signal properties that exhibit themselves as audible changes in pitch, loudness and syllable length ([20]). Prosodic features play a very important role in speech communication. For example, the use of *stress* can be used to provide *focus* to an element of discourse, perhaps highlighting new or important information. On the other hand, the *intonation* of an utterance, that is, the tonal (or melodic) aspects of its pitch[5], delivers information to the listener as to whether, for example, it is to be interpreted as a statement or as a question, or whether further information is to follow.

These prosodic aspects are part of the structure of a language and specific to it, in the same way that morphology and syntax are. Prosody can be studied at three levels ([20]):

Acoustic Level The acoustic manifestation of prosody (fundamental frequency, amplitude and duration) can be measured directly using specialized hardware or algorithms (such as pitch determination algorithms.)

Perceptual Level The perceptual level represents the prosodic events as heard by the average listener. The perceptual representation is accessible to the individual listener, but this mental representation can hardly be measured.

Linguistic Level The linguistic level represents the prosody of an utterance as a sequence of abstract units. The linguistic model is a structural interpretation of the data, which results from the analysis of prosodic data by a linguist.

3.2 Models for Prosody

A number of prosody transcription formalisms and methods have been developed over the years (both for TTS and as pure analysis techniques). Following from the above, they can also be classified as *acoustic*, *perceptual* or *linguistic* models, depending on the level at which the modelling is carried out.

3.2.1 Acoustic Models

Acoustic models focus on the prosodic aspects that emerge from the speech signal. Analysis and methods are thus carried out on F0 (fundamental frequency) contour curves, or representations thereof, using a number of parameters. One such is *Fujisaki's model*, which is based on the fundamental assumption that intonation curves, although continuous in time and frequency, originate in discrete events triggered by the reader, and that affect the fundamental frequency ([20]). Events are classified as phrase and accent commands, respectively modelled as pulses and step functions, which drive second order linear filters whose outputs are summed up to yield F0 values.

A more recent approach, the *Tilt* intonation model ([18]), views the intonation contour as a serious of pitch accent and boundary tone *events*. Events are modelled using three parameters—amplitude, duration and *tilt*, a dimensionless number that expresses the overall shape of the event.

3.2.2 Perceptual Models

Acoustic models can hardly be used for a linguistic, functional study of prosody, since the latter should at least take perception into account. Since the approximated F0 curves they generate are based on acoustic measures, they cannot guarantee that the details that are smoothed out are not really audible or that the ones that remain can really be heard.

Intonation models that operate at a *perceptual* level[6], on the other hand, aim to yield a quantitative but compact description of the prosodic attributes of the signal that are perceived. *Automatic perceptual stylisation* methods, for instance, base their analysis upon tonal segments identified from a stylised F0 contour. These perceived segments, however, might not be organized within a linguistic framework.

3.2.3 Linguistic Models

Linguistic models abstract the phonological nature of prosodic data beyond its acoustic or perceptual characteristics. This is a difficult task, since prosodic information is intrinsically relative (and does not correspond to absolute F0, duration or intensity values) and can be

[5] In some texts, the term *intonation* is used as a synonym for *prosody*.

[6] Here, the *perceptual* methods may be understood as a medium path between the sound signal oriented *acoustic* models, and the predominantly symbolic *linguistic* representations.

understood very differently by the listener, depending on the context. Laying linguistically oriented foundations of a transcription system for intonation requires a definition of both a prosodic vocabulary to account for the intonational events, and of a grammar that resolves the ambiguities that arise in the use of these symbols.

However, even though the *International Phonetic Alphabet* (IPA) ([35]) has been established for quite a while, no universally adopted prosodic syntax has been established to date, which is a fundamental lack for the development of linguistic theories and for the growth of speech applications. One can identify two main phonological models of intonation, which respectively analyse intonation in terms of *pitch contours* and *tone sequences*.

Pitch Contour Theory Under one school of thought, the intonation of an utterance originates as a sequence of elementary pitch contours taken from a limited set, which the speaker approximately realizes through his or her vocal apparatus. Each contour is seen as atomic, and cannot be decomposed into smaller parts. This theory led to the division of speech into *tone groups*, each composed of, at most, four components: *preheat*, *head*, *nucleus* and *tail*. The accented part of the tone group forms the nucleus, which is the only obligatory part and carries primary stress on its first syllable.

Tone Sequence Theory Unlike pitch contour theory, which considers absolute curves, tone sequence theory describes melodic curves in terms of relative tones. The seminal work in this respect is the treatise given for American English by Pierrehumbert ([50]), where tones are defined as the phonological abstractions for the target points obtained after an acoustic stylisation of the utterance.

Pierrehumbert distinguishes only two tones, a high tone (**H**) and a low tone (**L**) which are relatively contrasted: **H** is higher in the speaker's range than **L** would be in the same place. Sequences of **H** and **L** tones are restricted by a finite state grammar (Figure 3), which in turn distinguishes four categories of tones on the basis of distributional properties: *initial boundary tones*, *pitch accent tones*, *phrase accent tones* and *final boundary tones*. This theory has been more deeply formalized into the *Tones and Break Indices* (ToBI) ([57]) transcription system, which is further examined in Section 3.4.

3.3 Relating Prosody to Text

Prosody is related to text at different levels, including lexical, syntactical, semantic and pragmatic. A common misconception is that intonation is solely a function of the punctuation. Intuitively, it makes sense to assume that syntax plays an important role for prosody. Punctuation marks, however, are only a visible indicator of language structure of a language. Not all punctuation is prosodically relevant (as in the case, for example, of comma separated adjective lists) and not all prosodic phrases and boundaries will be marked by punctuation (another obvious example being sentence conjugation.)

Additionally, semantic and pragmatic context bear heavily upon the most appropriate prosodic interpretation of a text. One should consider, for instance, that given any arbitrary sentence, each individual word can be subject to *focus*, and hence to being stressed ([39, 20]), given different contexts.

Nevertheless, given the relative simplicity of syntactic analysis in comparison with the many problems raised by automatic handling of semantics and pragmatics, the possibility of describing prosody as a function of syntactic relationships between the words is of high relevance for TTS synthesis.

3.3.1 Syntactic Prosodic Parsing

The absence of a universal prosody transcription framework makes it hard to build a widely accepted theory of intonation. Moreover, having a scheme for notating the intonation of speech is only the starting point with respect to an implementation for TTS. The following step is to develop a mechanism for assigning such prosodic contours to new text, the main task of the *syntactic prosodic parser*. Dutoit ([20]) classifies syntactic-prosodic parsing techniques under three categories, namely *Heuristics Based*, *Grammar Based* and *Corpus Based* methods.

Hand-driven Heuristics Heuristics based mechanisms try to avoid the computational load and complexity of developing fully-fledged grammars and interpreting text by developing low-level heuristics that are able to uncover prosodic groups. The most basic of these is the organization of prosody on the basis of punctuation alone. As discussed above, such an approach is bound to have limited success in practice. A better approach is to take not only punctuation marks, but also lexical and grammatical words, as phrase delimiters. Liber-

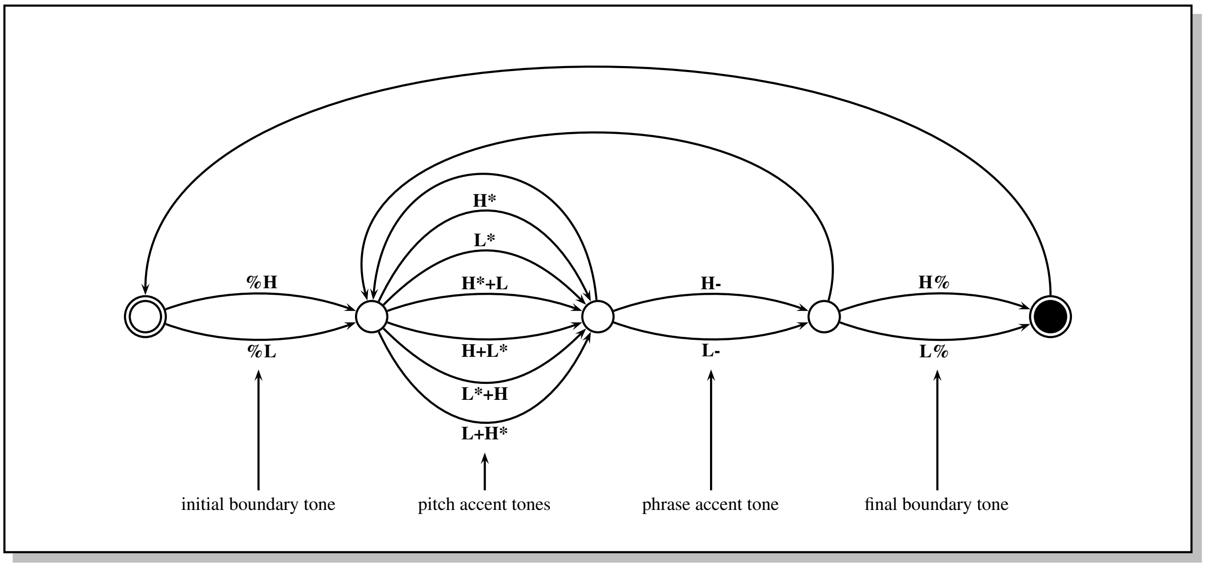

Figure 3: A Finite State Grammar for Tone Sequences

man and Church ([42]) report some success with a relatively crude algorithm, termed as the *chinks 'n chunks* algorithm, in which words are classified as being either *function* words or *content* words, and prosodic phrases are accounted for by a simple regular rule.

Heuristic based parsers have the advantage of being simpler to develop and generally faster in operation, as most of them attempt left-to-right, real-time analysis, instead of examining the entire sentence before producing parse results. They are also usually designed to be fail-safe in their operation. However, they rely heavily on the accuracy of part-of-speech tagging and errors in the latter can be lead to disastrous results. They also fail to account for higher-level linguistic phenomena, such as finding the beginning of a clause that does not start with a specific function word or the end of a clause that is not terminated by a comma.

Grammar Based Systems *Definite Clause Grammars* (DCGs) and *Unification Grammars* (UGs) offer a very flexible framework for real language processing. They implement context-free grammars augmented with variables, which naturally enables them to account for various kinds of agreement rules (such as gender, number and case) and to naturally transduce input sentences into their syntactic-prosodic representation by instantiating variables during the parsing process. (See the parser developed for German in [63] and also the use of feature structures for machine translation in [23].)

They easily lend themselves to *chart parsing*, a very powerful concept which allows the parser to record already parsed sub-structures into a table or a chart, so as to avoid analysing them again when backtracking. An advantage of chart parsing is that even if a full level parse is not achieved, it is still possible to select sub-optimal coverage by concatenating partial parses according to some criterion (such as by choosing the longest parses or by using a scoring mechanism.)

Once a syntactic structure has been derived, the gap remains to be bridged between prosodic and syntactic phrasing. Although the appropriate contours to apply may be embedded within the grammar rules and thus equated with the parse structure, it should be noted that the structure of the intonation phrases may in fact be "flatter" than that of the syntactic parse[7]. In order to cope with such mismatches, readjustment rules on the parse tree structure may be invoked, prior to selecting a corresponding prosodic contour. Grammar based systems have the advantage of supporting the embedding of linguistic insight within their rules, and hence can be incrementally refined with the support of a linguist expert. However, the inter-relationship between rules becomes increasingly more complex to manage as the rule set size becomes larger.

[7] Hence, in fact, the limited success of heuristic based approaches.

Automatic, Corpus-Based Methods Increasing the coverage of a set of rules (whether heuristic or not) often implies increasing the complexity of the analysis system, rendering the updating of a rule-based system in a manual fashion complex. Corpus based methods avoid that by using large text corpora, manually annotated with phrase boundaries and other prosodic information, and statistical inference algorithms in order to identify the rules from the text.

Techniques used for learning include *Neural Networks*, *Hidden Markov Models* and *Classification and Regression Trees* (CARTs.) With the latter approach, heuristics are organized into decision trees, where at each non-terminal node there is a yes or no answer about the value of the contextual factor associated with the node, and for each possible answer there is a branch leading to the next question. Efficient algorithms are available for building such trees from the training data set, having previously identified the appropriate factors that need to be addressed (in this case, syntactic context and the identification of phrase boundaries.)

Corpus based methods, and especially CART techniques, can produce good results, when proper training is carried out. However, the latter demands the availability of large, labelled databases, the availability of which is still lacking for Maltese. To attempt to develop such a corpus, one also assumes the existence of a standardized intonation description with which to annotate the utterance. This is in fact the subject of the Section 3.4.

3.3.2 Beyond Syntax

As already mentioned above, syntax is important for determining the default intonation of an utterance, but is not sufficient when it comes to selecting context sensitive contours. A number of formalisms, such as *Combinatory Categorial Grammars* ([51, 52]) and *Information Structures* ([41]), are being used in order to capture a deeper semantic interpretation of the discourse, and hence be in a better position to select contextually appropriate intonation.

A common idea is that of structuring the discourse into a *theme* (what the participants have agreed to talk about) and a *rheme* (what the speaker has to say about the theme.) The formalisms that are used to instantiate these structures can also have an implicitly assign intonational contour. In practice, however, studies are still limited to a certain class of restricted dialogue systems, such as expert system querying ([52]).

3.4 ToBI

As outlined in Section 3.2.3, *ToBI*, or Tones and Break Indices ([57]), is an intonation annotation framework that originated from Pierrehumbert's work ([50]) on the phonology and phonetics of English language. Whilst the original framework was defined for the English language, the concepts can in general be modified and applied to other languages. (See, for instance, [32] for German and [1] for Greek.)

The framework relies on describing speech on a number of parallel levels, referred to as "tiers." Figure 4 provides an illustration of a typical analysis, along with the F0 contour of the utterance ([6]). The most important tiers are the *tone tier*—which describes intonation by marking the F0 contour with **H** (High) and **L** (Low) marks, along with corresponding diacritics for marking accent and phrase boundaries—and the *break index tier*, which identifies well-formed intonational phrases and phrase boundaries with a numerical marker denoting an increasingly meaningful break in the utterance, starting from 0 (identifying the start of an utterance) to 4 (indicating its end). Additionally, there may be a *miscellaneous tier*, describing random events such as background noise or coughs. The descriptions are not meant to be universal for any language, but as guidelines on which specific details for each language variety may be specified.

3.4.1 Developing a ToBI Framework

What are the requirements for building a ToBI framework? Apart from the obvious need of a motivation, some guidelines are set for adopting ToBI to a new language variety ([62]). First of all, since a ToBI framework system is a community-wide standard, it requires a community of users who have agreed to adopt the conventions in database development and related research. It should also conform to a number of principles:

- The conventions should be accurate up to the current state of knowledge, and should be based upon a long-established body of research, especially as regards intonational phonology.
- An electronic recording of the original utterance is maintained along with the symbolic transcription.
- The conventions should be efficient, so that only meaningful pitch rises and falls are symbolically marked.

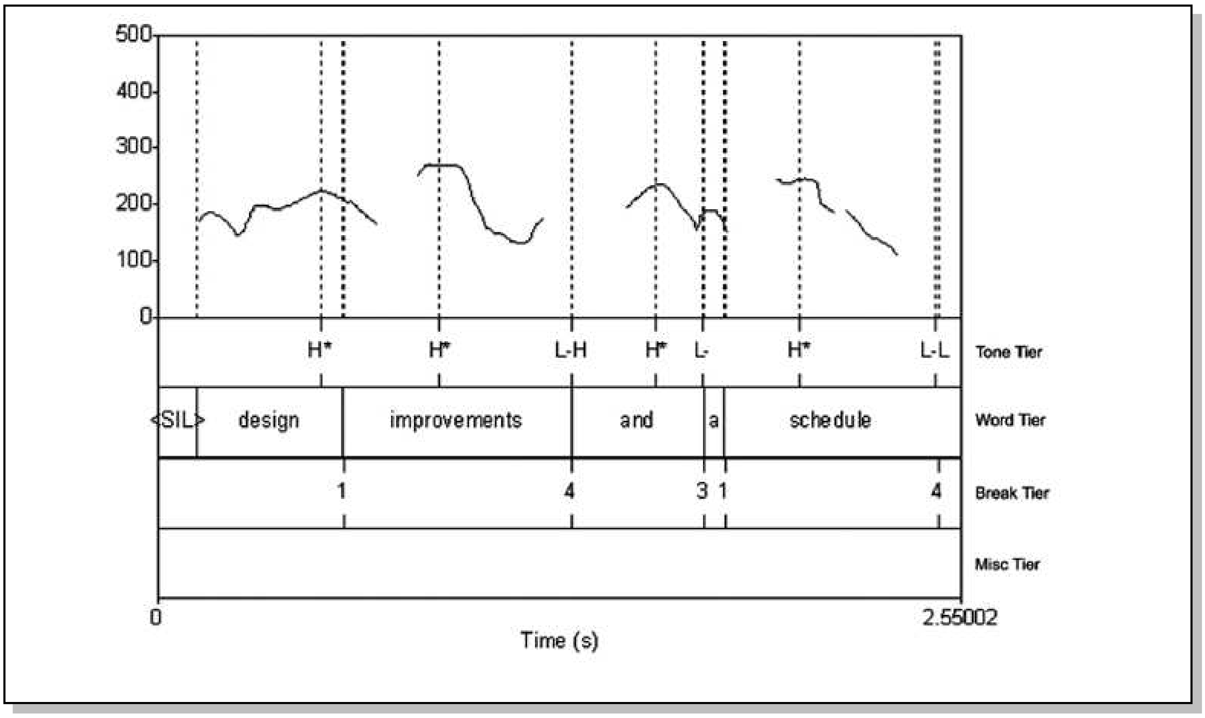

Figure 4: A Typical ToBI Analysis

- The conventions should be easy enough to teach to those who are not linguistic experts, and a manual is to be made available to new transcribers.
- Use of the conventions should be consistent, and development should include rigorous tests of inter-transcriber consistency.

A formalization of a ToBI annotation system for Maltese is hereby proposed as a reference framework for describing prosody for TTS. The motivation for this choice is given in the following section. It is noted here that a body of research for the intonational phonology of Maltese is already in existence ([3, 66, 11, 67]), while a community of linguistic and computational linguistics researchers that may benefit from such a framework is growing.

The remaining steps for defining a Maltese ToBI system would thus consist of the following:

- A formalization of the intonational descriptions available in previous research in a structured form.
- The establishment of a break index scheme.
- The compilation of annotation guidelines.
- The development and tagging of a speech database.
- The testing for inter-transcriber reliability.

3.4.2 Why ToBI?

Why adopt ToBI for Maltese prosody annotation and synthesis? It is not assumed to be a universal framework for intonation (in fact, it is expected that a different ToBI system is required for each language variety [62]) and may not necessarily be the best formalism for TTS—Black and Dusterhoff, for instance, report somewhat better synthesis results using an F0 contour generation scheme based on *Tilt* ([18]) than the one based on ToBI by Black and Hunt ([8]).

Nevertheless, ToBI is deemed a plausible choice for a number of reasons:

- It is a well-recognized framework within linguistic circles and possibly the most commonly used intonation description formalism in recent works.

- Its application for TTS has also been investigated, with a method for generating F0 contours from ToBI labels given in [8]. Support for ToBI is also including in the Festival TTS system ([28, 9]).

- Linguistic research addressing the intonation of Maltese and utilizing ToBI-style annotations has already been carried out in [66] and [67], where the intonation patterns that may be found in Maltese are extensively discussed. The availability of a ToBI based TTS synthesizer can then become a useful tool to the linguist for both validating and refining the study of Maltese intonation.

- It is independent of any prosodic assignment technique utilized, allowing the exploration of alternative methods within a well-designed modular TTS system, including the possibility of eventually introducing semantic-pragmatic analysis modules.

4 Project Resources

It should be clear, by now, that developing a TTS system is not a trivial process. It is natural, then, to make good use of all available resources that can simplify and expedite such an implementation. In the rest of this section, an overview is given of those software tools, data sets and research results that can be availed of for this purpose.

4.1 Software Tools

From Section 1, it follows that the TTS researcher needs to have an implementation of all the relevant modules in Figure 2 in order to obtain synthesized results, even if only interested in one aspect of the whole process. It is important, thus, to have the right tools that provide appropriate development frameworks, expedite the implementation of peripheral modules and allow testing of new and alternative ideas in a simplified fashion. A quick review, focused on freely available tools, is given below, organized hierarchically as shown in Figure 5.

4.1.1 TTS Frameworks

Festival is a general multilingual speech synthesis system developed at the *Centre for Speech Technology Research* (CSTR), University of Edinburgh ([28]). It offers a full TTS system, a number of voices and a set of APIs.

Flite (festival-lite) is a small, fast, run-time synthesis engine developed at *Carnegie Mellon University* (CMU) and primarily designed for small, embedded machines and/or large servers [25].

FreeTTS is a speech synthesis system written entirely in the Java programming language but based upon Flite ([26]).

MBROLA is a project initiated by the TCTS Lab of the *Faculté Polytechnique de Mons* (Belgium), the aim of which is to obtain a set of speech synthesizers for as many languages as possible, and provide them free for non-commercial applications ([46]).

4.1.2 Libraries and Support Tools

Festvox is another CSTR project ([29]) that aims to make the building of new synthetic voices more systemic, providing related documentation and useful scripts for both open and limited domain voices.

Edinburgh Speech Tools Library is a collection of C++ classes, functions and programs for manipulating the kind of objects used in speech processing ([22]). In particular, it provides utilities for building CARTs.

4.1.3 Acoustic Tools

Praat is a program that permits sophisticated analysis, synthesis, and manipulation of speech ([53]). It is particularly tailored for the linguist, and includes support for ToBI annotations.

SIL Speech Tools is a set of tools that facilitates phonetic transcription, acoustic analysis and phonological analysis [56].

Diphone Studio is a tool for developing and maintaining a set of diphones for use in speech synthesis ([17]).

EMU is a collection of software tools for the creation, manipulation and analysis of speech databases ([21]). Amongst other functions, it allows the researcher to organize speech segments based on the sequential and hierarchical structure of the utterances in which they occur.

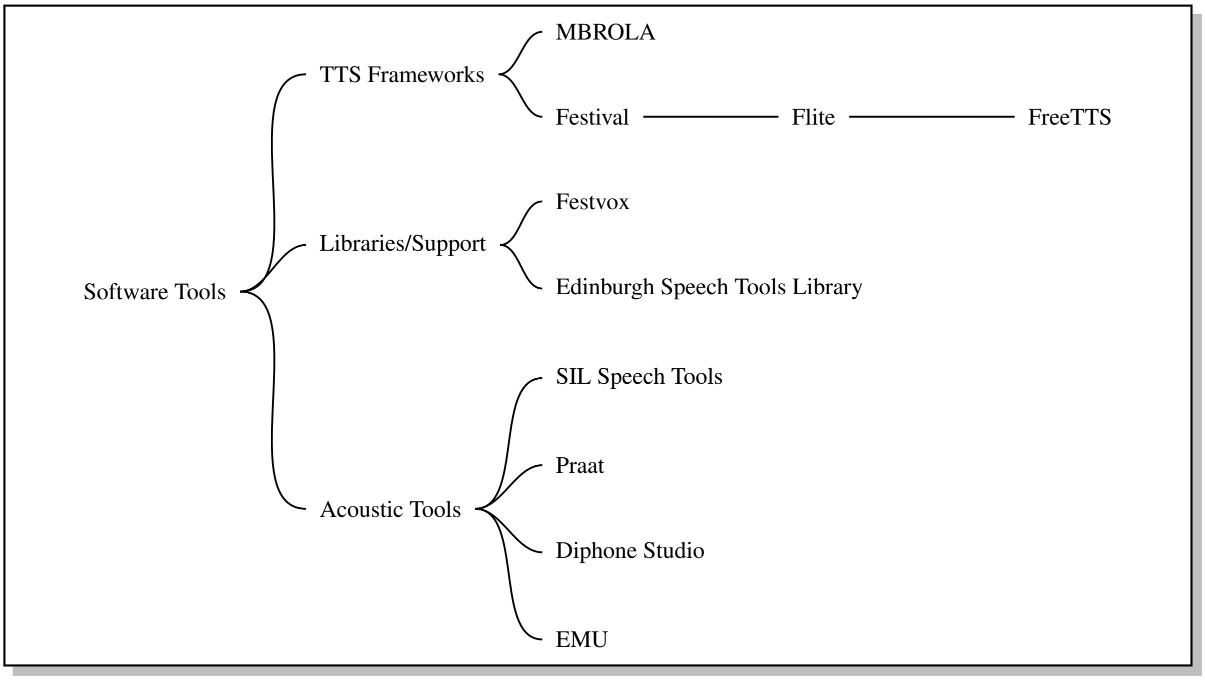

Figure 5: TTS Software Tools

4.2 Data Sets

A proper study of language intonation necessitates the use of an appropriately transcribed speech corpus in the target language. In order to include within such a corpus the linguistic aspects under investigation, the utterances elicited from multiple speakers are often constructed through the use of game tasks ([58, 33]) that implicitly impose upon the subjects the use of specific speech structures.

Such corpora, as, for example, the Darpa TIMIT database and the IViE corpus ([36]), are readily available for other languages. Unfortunately, no such corpus exists for the Maltese language. The recording of an appropriately designed set of utterances spoken by a mixed set of speakers is expected to be required for the formalization of a ToBI framework for Maltese[8].

A reasonably large lexicon is another required component for the project. Again, the availability of a suitable lexicon is an issue. Although Dalli ([16]) made some inroad on the development of a computational lexicon for Maltese, its completeness is still less than satisfactory. In the limit, a stochastically selected lexicon may be required for the purpose of developing a reference implementation.

In the case of diphone synthesis for TTS, which is the more commonly adopted synthesis approach in current systems, there is also the requirement of a recorded diphone database per speaker. For really natural results, it is necessary to record a native Maltese speaker, in order to capture the nuances of articulation. Building a diphone database, however, is a laborious and time-consuming effort, which is not a focal part of this project. The use of a diphone database for another language that captures the phonemic inventory of Maltese[9] may be sufficient for didactic purposes.

Finally, due to the selection of SMS as the input domain of focus, sets of such messages are obviously required for training and testing purposes. Such messages are available for the required analysis with the support of MobIsle Communications Ltd ([30]), whereby all header information, including sender number, receiver number and timestamps, are stripped off prior to usage, and only the message text is maintained. Messages are also selected at random, removing any implicit tem-

[8] In collaboration with Dr. Alexandra Vella, Faculty of Arts, University of Malta.

[9] An English diphone database, for instance, would probably capture all Maltese phonemes with the possible exception of the glottal plosive /ʔ/.

poral information. Furthermore, to guarantee complete anonymity, a filter is used to swap well-known person and place names with random ones prior to being selected for analysis.

4.3 Previous Research

In comparison with other languages, and English in particular, the area of TTS for Maltese has been little explored thus far, with only one other successfully implemented system ([47]). As regards Maltese phonology, a number of important references have already been mentioned ([3, 66, 11, 67]). Some previous projects have also tackled aspects of Maltese language processing which are of interest, including [48] (on statistical spell checking), [16] (on deriving a computational lexicon based on a phonemic transcription of the input) and more recently [12] (on the development of a voice recognition system for Maltese.) It is another objective of this project to support further linguistic insight into phonological aspects of the language by providing appropriate tools, whilst bringing to bear other such previous computational linguistic research.

5 Application

The input domain for the TTS system is set to be general SMS messages. In this section, this choice is motivated with an illustration of how such a system can be turned into a practical application.

In today's information age, there are a multitude of ways to communicate, including voice (fixed or mobile telephony), fax, email and instant messaging (including SMS). It is typical, especially for business professionals, to handle each type of message on a daily basis. With the increasing ubiquity of electronic mail and GSM devices, the time devoted to handling messages can be considerable. *Unified Messaging* ([65, 34]) is a vision that is concerned with the integration of the various message types into a central repository, so that a user can access them all in a standardized manner.

How could this work in practice? With Unified Messaging, a single point of access to all message types, be it voice, fax, email or SMS, is provided from virtually any communications device, whether from a mobile, or on a personal computer with a conventional Web browser. In the latter case, one could envisage the user's familiar email inbox being augmented with a unique icon that identifies each message type. The unified messaging system would take care of translating faxes into graphical images (or possibly, into text identified through OCR mechanisms) and voice mail messages into audio files.

When not on the desk or on the road, users can still access and manage all their messages through an IVR-based telephony user interface. Using this, one can dial into the unified messaging system from any telephone and be able to quickly and efficiently listen to and respond to any pending message. In this case, it would be the turn of email and SMS messages to be translated to speech in order to be communicated to the user, and hence the requirement of robust, high-quality TTS.

The maturity of speech application is expected to be useful to the telecommunications industry in other areas. One could imagine, in the not too distant future, a call-centre enabled or automated by speech applications ([40]) whereby a client can phone to an IVR or *Speech Recognition* system in order to lookup someone in the directory and get answered by a virtual agent that answers back through a *Concept to Speech* system. (A demo of such a prototype system can be found at [14].)

6 Research Aims

This project is intended to both deliver a usable end product, in the form of a robust text to speech synthesizer for SMS messages with mixed language support, and also to provide tools that can help further linguistic research on the Maltese language. Supported by the resources identified in Section 4, the aim is to develop a modular system that can handle our input domain and provide the means to establish a relationship between the syntactic structure of a given text and a corresponding intonational contour.

Essentially, an answer to the following question is searched for: "Given an arbitrary, non-contextualized, Maltese phrase, how can we identify the default intonational contour that a native speaker would apply?" The notion of *default* is stressed in formulating this question, given the input domain (from which structured context is difficult to extract) and the complexity of introducing semantic and pragmatic analysis. However, it is also intended to tailor developments in such a modular manner that allows the possibility of introducing the latter at a later stage (in the manner, for instance, of [41]).

A priori, a preference is given to focus on a grammar based system given that a tool developed on such an approach would allow the linguist to explicitly formulate or analyse the relationship between syntax and prosody, a facility which other approaches may not readily provide. In this respect, the aim is not to provide an all-encompassing theory for prosody assignment, but rather to develop a framework that a linguist can further fine-tune for better results.

6.1 Project Deliverables

In summary, at least the following components are identified as deliverables for this project:

- A framework for identifying code switching at the word level.
- A framework for handling incorrectly written special Maltese characters.
- A ToBI annotation system for Maltese.
- A syntactic-prosodic parser supporting a grammar description augmented with the ToBI notation description and capable of deriving an intonationally descriptive parse.
- An underlying framework for transforming the resulting parse into speech output. Festival/Flite/FreeTTS may be used for this purpose.
- A means for evaluating result qualities.

References

[1] Arvaniti, A., and Baltazani, M. Greek ToBI: A system for the annotation of Greek speech corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (Athens, 2000), vol. II, pp. 555–562.

[2] AT&T labs-research interactive multi-lingual demo. http://www.research.att.com/projects/tts/demo.html.

[3] Azzopardi, M. *The Phonetics of Maltese: Some Areas Relevant to the Deaf*. PhD thesis, University of Edinburgh, 1981.

[4] Baart, J. L. G. *A Field Manual of Acoustic Phonetics*. SIL International, July 2001.

[5] Bailly, G., and Benoit, C., Eds. *Talking Machines*. North-Holland Elsevier Science Publishers B.V., Sara Burgerhartstraat 25, P.O. Box 211, 1000 AE Amsterdam, The Netherlands, 1992.

[6] Beckman, M. E., and Elam, G. A. *Guidelines for ToBI Labelling*, 3.0 ed. The Ohio State University Research Foundation, Mar. 1997.

[7] Beesley, K. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Languages at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association* (12–16 Oct. 1988), pp. 47–54.

[8] Black, A. W., and Hunt, A. J. Generating F0 contours from ToBI labels using linear regression. In *Proceedings of ICSLP 96* (1996), vol. 3, pp. 1385–1388.

[9] Black, A. W., Taylor, P., and Caley, R. *The Festival Speech Synthesis System*, 1.4 ed. University of Edinburgh, 27 Dec. 2002.

[10] Borg, A. *Ilsienna*. Albert Borg, 1988.

[11] Borg, A., and Azzopardi-Alexander, M. *Maltese*. Descriptive Grammars. Routlege, London and New York, 1997.

[12] Calleja, S. Maltese speech recognition over mobile telephony. In Cordina and Pace [15].

[13] Cavnar, W. B., and Trenkle, J. M. N-gram-based text categorization. In *Proceedings of (SDAIR)-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, US, 1994), pp. 161–175.

[14] Cmu communicator. http://fife.speech.cs.cmu.edu/Communicator.

[15] Cordina, J., and Pace, G. J., Eds. *CSAW '03, Computer Science Annual Workshop* (10, 17 July 2003), University of Malta, Computer Science Department.

[16] Dalli, A. Computational lexicon for Maltese. B.Sc. IT final year dissertation, University of Malta, Apr. 2002.

[17] Diphone studio homepage. http://www.fluency.nl/dstudio/dstudio.htm.

[18] DUSTERHOFF, K., AND BLACK, A. W. Generating F0 contours for speech synthesis using the Tilt intonation theory. In *Proceedings of ESCA Workshop of Intonation* (Athens, Greece, Sept. 1997), pp. 107–110.

[19] DUTOIT, T. High-quality text-to-speech synthesis: An overview. *Journal of Electrical and Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis 17*, 1 (1997), 25–37.

[20] DUTOIT, T. *An Introduction to Text-To-Speech Synthesis*, vol. 3 of *Text, Speech and Language Technology*. Kluwer Academic Publishers, P.O. Box 322, 3300 AH Dordrecht, The Netherlands, 1997.

[21] Emu homepage. http://emu.sourceforge.net.

[22] Edinburgh speech tools homepage. http://www.cstr.ed.ac.uk/projects/speech_tools.

[23] FARRUGIA, P.-J. An automatic machine translation system for English-Maltese. B.Sc. IT final year dissertation, University of Malta, 1999.

[24] FARRUGIA, P.-J. Text-to-speech technologies for mobile telephony services. In Cordina and Pace [15].

[25] Flite homepage. http://www.speech.cs.cmu.edu/flite/index.html.

[26] Freetts homepage. http://freetts.sourceforge.net.

[27] FURUI, S., AND SONDHI, M. M., Eds. *Advances in Speech Signal Processing*. Marcel Dekker, Inc., 270 Madison Avenue, New York, New York 10016, 1991.

[28] Festival homepage. http://www.cstr.ed.ac.uk/projects/festival.

[29] Festvox homepage. http://www.festvox.org.

[30] MobIsle Communications Ltd. http://www.go.com.mt.

[31] GREFENSTETTE, G. Comparing two language identification schemes. In *3rd International Conference on Statistical Analysis of Textual Data* (Rome, 11–13 Dec. 1995).

[32] GRICE, M., BAUMANN, S., AND BENZMÜLLER, R. German intonation in autosegmental-metrical phonology. In Jun [38].

[33] Hcrc map task corpus. http://www.hcrc.ed.ac.uk/maptask.

[34] IEC unified messaging tutorial. http://www.iec.org/online/tutorials/unified_mess.

[35] IPA homepage. http://www2.arts.gla.ac.uk/IPA/ipa.html.

[36] IViE corpus. http://www.phon.ox.ac.uk/~esther/ivyweb.

[37] Java speech api homepage. http://java.sun.com/products/java-media/speech.

[38] JUN, S.-A., Ed. *Prosodic Typology: Through Intonational Phonology and Transcription*. Oxford University Press, (in press).

[39] JURAFSKY, D., AND MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Prentice-Hall, Inc., Pearson Higher Education, Upper Saddle River, New Jersey 07458, 2000.

[40] KANELLOS, M. Talking computers nearing reality. http://news.com.com/2100-1008-1023966.html?part=dht&tag=ntop, July 2003.

[41] KRUIJFF-KORBAYOVÁ, I., ERICSSON, S., RODRÍGUEZ, K. J., AND KARAGJOSOVA, E. Producing contextually appropriate intonation in an information-state based dialogue system. In *ICPhS* (12–17 Apr. 2003), pp. 227–234.

[42] LIBERMAN, M. Y., AND CHURCH, K. W. Text analysis and word pronunciation in text-to-speech synthesis. In Furui and Sondhi [27], ch. 24, pp. 791–831.

[43] LINGGARD, R. *Electronic Synthesis of Speech*. Cambridge University Press, 1985.

[44] Loquendo vocal technologies and services. http://www.loquendo.com.

[45] MARY TTS system homepage. http://mary.dfki.de.

[46] Mbrola homepage. http://tcts.fpms.ac.be/synthesis/mbrola.html.

[47] MICALLEF, P. *A Text To Speech System for Maltese*. PhD thesis, University of Surrey, 1998. unpublished.

[48] MIZZI, R. The development of a statistical spell checker for Maltese. B.Sc. IT final year dissertation, University of Malta, June 2000.

[49] Maltilex homepage. http://mlex.cs.um.edu.mt.

[50] PIERREHUMBERT, J. B. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, September 1980.

[51] PREVOST, S., AND STEEDMAN, M. Generating contextually appropriate intonation. In *Proceedings of the Sixth Conference of the European Chapter of ACL* (Utrecht, 1993), pp. 332–340.

[52] PREVOST, S., AND STEEDMAN, M. Specifying intonation from context for speech synthesis. *Speech Communication 15*, 1–2 (1994), 139–153.

[53] Praat homepage. http://www.praat.org.

[54] SANTEN, J. P. V., SPROAT, R. W., OLIVE, J. P., AND HIRSCHBERG, J., Eds. *Progress in Speech Synthesis*. Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA, 1996.

[55] Scansoft homepage. http://www.scansoft.com.

[56] Sil speech tools homepage. http://www.sil.org/computing/speechtools.

[57] SILVERMAN, K., BECKMAN, M., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C., PRICE, P., PIERREHUMBERT, J., AND HIRSCHBERG, J. ToBI: A standard for labeling English prosody. In *ICSLP-92* (1992), vol. 2, pp. 867–870.

[58] SPEER, S., WARREN, P., AND SCHAFER, A. Intonation and sentence processing. *Symposium Paper at International Congress of Phonetic Sciences* (2003).

[59] SPROAT, R., Ed. *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachussets 02061, USA, 1998.

[60] SPROAT, R., HUNT, A., OSTENDORF, M., TAYLOR, P., BLACK, A., LENZO, K., AND EDGINTON, M. Sable: A standard for tts markup. In *Proceedings of ICSLP 98* (1998), R. H. Mannell and J. Robert-Ribes, Eds., vol. 5.

[61] SPROAT, R., TAYLOR, P., TANENBLATT, M., AND ISARD, A. A markup language for text-to-speech synthesis. In *Proceedings of EUROSPEECH 97* (Rhodes, Greece, 1997).

[62] ToBI homepage. http://www.ling.ohio-state.edu/~tobi.

[63] TRABER, C. *SVOX: The Implementation of a Text-to-Speech System for German*. PhD thesis, Swiss Federal Institute of Technology, 1995.

[64] Unicode homepage. http://www.unicode.org.

[65] Unified messaging news and information portal. http://www.unifiedmessaging.com.

[66] VELLA, A. *Prosodic Structure and Intonation in Maltese and its Influence on Maltese English*. PhD thesis, University of Edinburgh, 1995.

[67] VELLA, A. Phrase accents in Maltese: distribution and realisation. In *ICPhS* (2003).

[68] Voicexml homepage. http://www.voicexml.org.