Language Resources and Evaluation for the Support of the Greek Language in the MARY TtS

Pepi Stavropoulou[1,2], Dimitrios Tsonos[1], and Georgios Kouroupetroglou[1]

[1] National and Kapodistrian University of Athens,
Department of Informatics and Telecommunications, Greece
{pepis, dtsonos, koupe}@di.uoa.gr
[2] University of Ioannina, Department of Philology, Greece

Abstract. The paper outlines the process of creating a new voice in the MARY Text-to-Speech Platform, evaluating and proposing extensions on the existing tools and methodology. It particularly focuses on the development of the phoneme set, the Grapheme to Phone (GtP) conversion module and the subsequent process for generating a corpus for building the new voice. The work presented in this paper was carried out as part of the process for the support of the Greek Language in the MARY TtS system, however the outlined methodology should be applicable for other languages as well.

Keywords: MaryTTS, Greek Language, Grapheme to Phone, Diphone Database

1 Introduction

Among the most important factors for determining the success of a Text to Speech system is the quality of the lexicon, the Grapheme to Phone (GtP) module, and most importantly the quality of the speech database. The latter is conditioned on the set of phonemes used for defining the pronunciations and consequently the set of diphones for the TtS language, as well as the corpora and methods used for selecting the final set of utterances to be recorded by the voice talent. The final set should ideally provide maximum coverage of the possible diphones in a language and also accommodate important linguistic and prosodic events, such as question or negation contours.

The ultimate goal is the generation of a corpus that can effectively and adequately support the development of a more "context-aware" voice, which could be used in contexts such as natural language spoken dialogue systems, companion robots and so forth. Most generic TtS systems are trained on neutral, read speech databases, which both differ in style and lack pragmatic, "context-sensitive" events often occurring in dialogue [1]. Dialogue, on the other hand, poses specific requirements on the corpus used for developing the TtS voice. At minimum there should be an adequate representation of different types of questions, list structures, and deaccented materials as a result of early, non-default focus position [2,3].

As a first step we used the existing tools and processes for building a new language in the MARY TtS platform [4]. More specifically, in this paper we describe and evaluate the tools and processes for building the speech database, so as to ultimately extend them in the line of thought described above.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 523–528, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

2 The MARY TTS – New Voice Support Process

The MARY TtS system follows the Client-Server (CS) model. Server side executes the: text preprocessing/normalization, natural language processing, calculation of acoustic parameters and synthesis. The client sends the server the requests including the text to be processed and the parameters for the text handling by the server side. The system is: multi-threaded, due to CS implementation, flexible (modular architecture) and transparent and understandable as possible using XML-based, state-of-the-art technologies such as DOM and XSLT [4]. MARY TtS also includes several tools, in order to easily add support for a new language and build Unit Selection and HMM-based synthesis voices [5,6]. The tools can be grouped into: a) New Language Support Tools (NLST) and b) Voice Creation Tools (VCT). Using NLST we are able to create a new language support, providing the minimum NLP component for MARY TtS, and a text corpus in order to support the next stage using VCT, for the implementation of Unit Selection or HMM-based voice.

Greek language currently is not supported in MARY TtS. Following are the basic, necessary steps for baseline support of a new language in the MARY framework [5]:

- Define the set of allophones for the new language.
- Build the pronunciation lexicon, train Letter to Sound rules (i.e. GtP module) based on handcrafted transcriptions and define a list of functional words for the development of a primitive POS tagger.
- Prepare recording script, using Wikipedia corpus in combination with provided tools.
- Record the prompts, for the speech corpus creation based on Wikipedia corpus.
- Run Voice Import tools to build a Unit Selection or HMM-based voice.

3 Allophone Set Definition

The first step for developing a new language is the definition of the language's allophone set. The set is then used for the development of the grapheme to phone module and the generation of the diphone database. In general, the set of allophones should provide an adequate representation of the phonemic structure of the language, while at the same time the number of phones is kept manageable. Standard principles for deriving the set of allophones are the minimal pair distinction principle (cf. e.g. [7]) and phonetic similarity [8]. The latter ensures that same half phone units within diphones join well together.

Accordingly, the set defined for the Greek language was based on the state of the art descriptions of the language's phonemic inventory [9]. The selection of allophones was further conditioned on their systematic – and hence predictable – behavior which also allowed for a consistent, standardized representation in the lexicon. In this line of thought, within word freely alternating allophones such as [mb] and [b] were denoted by a single abstract phone. The choice between the two pronunciations is often speaker dependent subject to characteristics such as speaker's age, origin and dialect. On a later step, this abstract pronunciation could be specified to meet individual's speakers idiolect. On the other hand, mutually exclusive allophones such as [k] and [c] had distinct representations. Approximants [ts] and [tz] were represented as a single phoneme, contrary to [ks] and [ps], which correspond to two distinct phones and are thus not represented as a single unit in the phones set. Finally, contrary to previous representations for the Greek Language [10], no distinction was made between stressed and unstressed vowels in the inventory (i.e. no distinct allophones). We thus reduced the size of the inventory and represented stress as an abstract separate linguistic entity instead, affecting the complete syllable and consequent unit selection. Table 1 illustrates the final set of allophones used.

Table 1. Greek Allophone Set

Symbol	Example Word	Symbol	Example Word
a	anemos “wind”	G	Gala “milk”
e	eTnos “nation”	J	jenos “origin”
i	isos “maybe”	x	xara “joy”
o	oli “all”	C	Ceri “hand”
u	urios “favourable”	m	moni “alone”
p	poli “city”	M	Mazo “resemble”
b	bala “ball”	n	neos “new”
t	telos “end”	N	NoTo “feel”
d	dino “dress”	r	roi “flow”
k	kozmos “world”	R	tReno “train”
c	cima “wave”	l	lemoni “lemon”
g	gol “goal”	L	Lono “melt”
q	qiNa “bad luck”	ts	tsiGaro “cigarette”
f	filos “friend”	dz	dzami “glass”
v	velos “arrow”	W	laWo “glow”
T	Telo “want”	V	meVa “mint”
D	Dino “give”	Y	aYaLa “hug”
s	stelno “send”	Q	aQizo “touch”
z	zoi “life”	-	-

4 Grapheme to Phone Conversion

The Grapheme to Phone (GtP) module generates the pronunciation of words based on graphemes, the letters they are comprised of. Typically the word is first looked up in the pronunciation lexicon and if not found, its pronunciation is automatically generated by the GtP module. In the Open Mary platform the grapheme to phone rules are automatically learnt from phonetically transcribed data. The technique is based on CART decision trees that use questions on each grapheme’s context (i.e. preceding and subsequent letters/graphemes). In our case, questions based on a context of two graphemes both subsequent and following turned out to yield the best results.

A set of 4900 words was initially used for learning the GtP rules. 3242 words and their corresponding pronunciation were automatically extracted from an annotated, transcribed set of 600 newspaper sentences, which were developed as part of the RHETOR project [11]. The rest of the words were manually added, to account for grapheme sequences that did not occur in the original set and corresponded to various phonological rules affecting pronunciation. Furthermore, transcriptions were enriched to include syllable structure and stress patterns. Syllabification adhered to the maximum onset principle. Following is a typical entry from the lexicon: *πληϱοφοϱική* | p l i - r o - f o - r i - ’ci

The current model had a 5,2% word error rate on a test set consisting of 500 frequently occurring Greek words. Word frequencies were based on statistics of use from the Greek Wikipedia repository. To improve the model’s performance we reduced the “minLeafData” option to 35 from 100 which was the default value. The “minLeafData” value determines the minimum number of instances that have to occur in at least two subsets induced by split, i.e. at a leaf node. 35 proved to be a safe threshold, given that Greek has a rather regular orthography. In general, though, in languages such as e.g. Greek, German or Spanish, where the grapheme to phone relationship is highly structured and the orthography regular, rule based approaches can be just as – or even more – effective as data driven techniques.

5 Recording Script Preparation – Diphone and Intonation Coverage

The Greek Wikipedia repository [12] and the basic NLP module of MARY TtS were used for the recording script preparation. The size of Wikipedia xml file was 824.5MB. The file was processed by the sentence cleanup procedure, and stripped-off from any annotation/xml tag, keeping only the text content. Next, an automated cleanup procedure of Wikipedia content was executed, in order to exclude sentences with unknown words or strange symbols and to extract only reliable sentences with the optimum diphone coverage.

Of course, the automated procedure cannot completely remove all unreliable sentences, thus a manual selection is mandatory. We excluded sentences, which are missed by the automated procedure, containing e.g. duplicate sentence entries, any special/unknown characters, non-Greek words/characters, difficult to be read aloud by the voice talent during voice recording.

The tools for the recording script preparation provide statistical results at the end of each process. Table 2 presents the initial and final corpus diphone coverage. Initial corpus consists of 13,876 sentences, covering a total number of 1,990,345 diphones and 834 different diphones. It should be noted that not all diphones are truly possible given language specific phonotactic rules. The final manually selected set is comprised of 1,243 sentences and drastically reduced diphone coverage (524 different diphones).

With regards to the coverage of different sentence types (i.e. questions, negation, list structures), affirmative declaratives comprised the vast majority of the selected sentences, while wh-questions, polar questions and negation comprised a mere 3.18%, 1.62% and 7.78% of the total sentences respectively. Accordingly, diphone coverage

Table 2. Initial diphone coverage distribution and corresponding number of sentences for the Wikipedia Corpus before and after manual selection respectively.

	Initial Coverage	Manual Selection
Number of Sentences	13,876	1,243
Diphone Coverage	834	524

in these contexts was significantly limited, which is expected to have a negative effect on the final output of the synthesizer, especially in the case of unit selection synthesis whereas there is often limited signal processing modification. Furthermore, at this point the simple intonation coverage algorithm makes no distinction between different types of questions or differences in polarity. Nevertheless, in many languages, e.g. English and Greek among others, wh-questions and polar questions have distinct phonological melodies; it is a fundamental distinction that should therefore be modeled. We recorded the final sentence set (approximately 90 minutes of recordings, 16 bit – 44.1 KHz sampling rate) in order to create a speech corpus for the later voice import procedure. Voice import tools (for a detailed description see [13]) were executed, changing the sampling rate to 16 KHz, with their default configurations and a Unit Selection Voice and a HMM-based Voice were created.

6 Conclusions

In conclusion, diphone coverage seems to be rather low, especially after manual selection. The number of distinct diphones drops to 524 which is almost half the number achieved in e.g. [14]. In their study [14], the construction of a database for a Greek TtS system is presented, which achieves coverage for 813 diphones. The initial corpus included 300 most frequent diphones. They inserted new sentences containing the missing diphones, achieving maximum coverage. In our case, since manual selection is a post process, it is not accounted for by the greedy algorithm used for the initial utterance selection.

Furthermore, only the basic intonation coverage is achieved. Wikipedia does not seem to be the optimal resource for ensuring a sufficient coverage for context-aware dialogue based applications. It is notable that most instances of questions occurred in the comments section rather than the main article. In addition, the simple "punctuation based" features used in the greedy algorithm should be enriched with additional features (e.g. wh-words, negation particles) to identify different types of sentences corresponding to different prosodic realizations. The selection process would be more complete with the use of additional features such as utterance position, position relative to the operator (wh-word, negation particle) and so forth, always taking into account the issues of data sparsity and inventory size.

Accordingly, we need to explore other types of databases containing actual transcribed dialogues (examples are e.g. the Switchboard corpus for English or the Corpus of Greek Text [15] for Greek). Nevertheless, human-human dialogues may not be as restricted and still lack events that are particular to the “human-machine dialogues” genre (e.g. lists). Therefore, we may need to resort to a hybrid approach where the “normal”, real life text material is supplemented with hand crafted material, designed to ensure sufficient coverage.

7 Acknowledgments

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) under the Research Funding Project: “THALIS-University of Macedonia-KAIKOS: Audio and Tactile Access to Knowledge for Individuals with Visual Impairments”, MIS 380442.

References

1. Syrdal, A., and Kim, Y-J.: Dialog speech acts and prosody: Considerations for TTS. In Proc. of the Speech prosody, Brazil (2008)
2. Huang, X., Acero, A., and Hon, H.W.: Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR (2001)
3. Stavropoulou, P., Spiliotopoulos, D., and Kouroupetroglou, G.: Where Greek Text to Speech Fails. In Proc. of the 11th International Conference on Greek Linguistics, Rhodes, Sept. 2013.
4. Schröder, M., and Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology 6-4, 365-377 (2003)
5. Pammi, S., Charfuelan, M., and Schröder, M.: Multilingual Voice Creation Toolkit for the MARY TTS Platform. LREC 2010, Malta (2010)
6. Schröder, M., Charfuelan, M., Pammi, S., and Steiner, I.: Open source voice creation toolkit for the MARY TTS Platform. Proc. Interspeech. Florence, Italy (2011)
7. Ladefoged, P., Johnson, K.: A Course in Phonetics. Wadsworth, Cengage Learning Inc., Boston (2010)
8. Taylor, P.: Text to Speech Synthesis. Cambridge University Press, Cambridge (2009)
9. Arvaniti, A.: Greek Phonetics: The State of the Art. Journal of Greek Linguistics 8: 97-208 (2007)
10. Fotinea, S.-E. and Tambouratzis, G.: A Methodology for Creating a Segment Inventory for Greek Time Domain Speech Synthesis. International Journal of Speech Technology, Vol. 8, No. 2, pp. 161-172 (2005)
11. Fourli-Kartsouni, F., Slavakis, K., Kouroupetroglou, G. and Theodoridis, S.: A Bayesian Network Approach to Semantic Labelling of Text Formatting in XML Corpora of Documents. Lecture Notes in Computer Science, 4556: 299-308 (2007)
12. Wikipedia: http://dumps.wikimedia.org/elwiki/latest/elwiki-latest-pages-articles.xml.bz2, accessed, May (2013)
13. Voice Import Tools Tutorial : How to build a new Voice with Voice Import Tools: http://mary.opendfki.de/wiki/VoiceImportToolsTutorial
14. Fotinea, S.-E., Tambouratzis, G., and Carayannis, G.: Constructing a segment database for greek time domain speech synthesis. Proc. of EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, 3-7 September, Aalborg, Denmark (2001)
15. Corpus of Greek Text: http://www.sek.edu.gr