ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication xxx (2006) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Tone-Group F_0 selection for modeling focus prominence in small-footprint speech synthesis

Gerasimos Xydas, Georgios Kouroupetroglou *

University of Athens, Department of Informatics and Telecommunications, Division of Communication and Signal Processing, Panepistimiopolis, Ilisia, GR-15784 Athens, Greece

Received 13 July 2005; received in revised form 28 December 2005; accepted 1 February 2006

Abstract

This work targets to improve the naturalness of synthetic intonational contours in Text-to-Speech synthesis through the provision of prominence, which is a major expression of human speech. Focusing on the tonal dimension of emphasis, we present a robust unit-selection methodology for generating realistic F_0 curves in cases where focus prominence is required. The proposed approach is based on selecting Tone-Group units from commonly used prosodic corpora that are automatically transcribed as patterns of syllables. In contrast to related approaches, patterns represent only the most perceivable sections of the sampled curves and are encoded to serve morphologically different sequence of syllables. This results in a minimization of the required amount of units so as to achieve sufficient coverage within the database. Nevertheless, this optimization enables the application of high-quality F_0 generation to small-footprint text-to-speech synthesis. For generic F_0 selection we query the database based on sequences of ToBI labels, though other intonational frameworks can be used as well. To realize focus prominence on specific Tone-Groups the selection also incorporates a level indicator of emphasis. We set up a series of listening tests by exploiting a database built from a 482-utterance corpus, which featured partially purpose-uttered emphasis. The results showed a clear subjective preference of the proposed model against a linear regression one in 75% of the cases when used in generic synthesis. Furthermore, this model provided ambiguous percept of emphasis in an experiment featuring major and minor degrees of prominence.

Keywords: Text-to-speech synthesis; Tone-Group unit-selection; Intonation and emphasis in speech synthesis

1. Introduction

Emphasis is essentially the use of language that humans employ in order to bring to prominence selective parts of speech and mainly convey non-lexical and pragmatic information. It primarily signals contrast (contrastive focus), distinction between new and given information (focus as the missing variable in a proposition), meaning pronunciation and mood or other emotions. Generally, it points

Abbreviations: HRG, Heterogeneous Relation Graph; LR, Linear Regression; MPE, Mean Perceived Emphasis; NLP, Natural Language Processing; TI, Tone Item; TG, Tone Group; TGS, Tone-Group Selection; TtS, Text-to-Speech.

* Corresponding author. Address: University of Athens, Efkalypton 39, Agia Paraskevi, GR-15342 Athens, Greece. Tel.: +30 2107275305; fax: +30 2106018677.

E-mail addresses: gxydas@di.uoa.gr (G. Xydas), koupe@di.uoa.gr (G. Kouroupetroglou).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.02.002

out the most important parts in an utterance. Humans use a collection of different prosodic aspects to denote emphasis when they speak. The most common are pause insertions before and after the emphasized words, duration stretching, intensity, and substantial pitch rate change. The latter has proven to be the most significant factor for the perception of prosody ('t Hart et al., 1990; d'Alessandro and Mertens, 1995; Xub and Sun, 2002).

Human speech communication is emphasized by its nature. Most sentences have at least one focus and this is something that is partially ignored in most prosody modeling works, providing "neutral" or "generic" coverage in preliminary prototypes; however this is not the case in real speech. One of the drawbacks of Text-to-Speech (TtS) synthesis that leads to monotonous prosodic cues is the lack of focus prominence over the corresponding segments of speech. Therefore, emphasis modeling and provision is a mean to increase the expressiveness and thus naturalness of synthetic speech.

A TtS synthesis system mainly consists of two components (Dutoit, 1997; Sproat, 1998): the natural language processing (NLP) and the signal processing. The first one deals with the text-to-prosody part, providing the latter with sufficient segmental and prosodic information to generate an appropriate acoustic signal that "resembles human speech well enough for the human brain to interpret it as such" (Clark, 2003). The generation of the prosodic structure is derived in the synthesis chain from higher-level linguistic analysis of utterances carried by the NLP component. To represent this specification, several intonational frameworks have been proposed by linguists as well as engineers, ranging from qualitative (e.g. ToBI (Silverman et al., 1992)) to quantitative (e.g. Tilt (Taylor, 2000; Dusterhoff and Black, 1997)). They model intonation in terms of segmental anchoring and type, as for example, which syllables deserve a pitch accent and what value, type or shape should that accent be of. To incorporate this intonational description in the acoustic signal, the F_0 modeling component generates a continuous pitch curve from these events (location and type of accent). The resultant degree of naturalness of the synthetic pitch is closely related to the quality of the events. F_0 modeling is of great importance in any signal processing approach, from formant synthesis (defining the F_0 parameter) and diphone-based concatenative synthesis (defining pitch modifications) to unit-selection synthesis, as prosody selection is also of significant factor in the latter (Campbell, 1994).

The rule-based F_0 generation approaches have given place to machine learning ones. The most commonly used statistical method is the Linear Regression (LR) (Black and Hunt, 1996). This offers reasonable pitch generation, especially when the input conditions match the training ones. Objective evaluations have reported correlation between the training and the observed data from 0.6 to 0.8 in generic conditions (Black and Hunt, 1996; Xydas et al., 2005). On the other hand, subjective experiments usually contrast with the good statistical results, as prosody is usually judged as adequate but rarely natural. The modified suprasegmental structure of utterances and the lack of prominence seem to affect the naturalness of the delivered prosody, as well as the normalization of timing during pitch alignment. To overcome this problem, corpus-based F_0 models have been proposed and the recent related research focuses on optimizing (a) model's design in order to achieve adequate data coverage within reasonably sized databases (Black and Lenzo, 2003; Schweitzer et al., 2003) and (b) selection algorithms that not only minimize joining costs but also reveal the semantics of prosody (Bulyko and Ostendorf, 2001; Quazza et al., 2001; Wightman et al., 2000).

1.1. Corpus-based F_0 modeling

Following the natural effects on the segmental quality of corpus-based speech synthesis (Hunt and Black, 1996), corpus-based F_0 modeling (Huang et al., 1996; Malfrere et al., 1998; Meron, 2001; Raux and Black, 2003) attempts to maintain the suprasegmental structure intact thus achieving finest tonal representation. The minimization of the concatenation cost between jointed units affects the overall smoothness of the contour, based on the available inventory and the selection algorithm. However, natural curves are preserved at least over the range of each selected unit. In each case, the delivered speaking style originates and hardly deviates from that of the original human speaker.

In (Huang et al., 1996) the intonation cues of a group of consecutive syllables that form a clause, constitute an F_0 template. The template database is constructed in such a way so that it includes only one instance of each template. In (Malfrere et al., 1998), a sequence of successive words ending in a content word forms a pattern (intonational group).

Selection is based on units with identical pattern to the target one; thus only equal number of syllables and events make a match. In (Meron, 2001), the selection algorithm goes down to the syllable level. A prosodic unit inventory is utilized to select units that match a set of target events. A Viterbi-like dynamic programming method is used to find a path with minimum cost. Units are selected with no restrictions, even in mid-word positions whereas F_0-smoothing is performed to deal with F_0 discontinuities. Also, that work targets to duration selection as well, and thus only candidates with matching number of syllables to the target units are considered, in order to allow the proper duration alignment during synthesis. A similar method was followed in (Raux and Black, 2003) where the size of F_0 selection units was reduced to single segments to allow more flexible modeling, whereas the evaluation showed that the selection was mostly performed on syllable boundaries. These approaches have to face the major problem of coverage within the database, as missing units might cause either bad selection or selection from smaller units if available. However, the latter nullify the purposes of the tonal selection, as the semantic and pragmatic information that prosody carries is difficult to be chunked in small pieces. This problem is even more important in cases of small-footprint speech synthesis, where the limited resources do not allow the selection from a large inventory.

1.2. Unit size

The definition of the appropriate intonational unit size differs from the spectral unit-selection case, where the preferred unit size ranges from words (Black and Lenzo, 2000a), to syllables, diphones (Conkie and Isard, 1994), phones (Black and Lenzo, 2000b), half-phones (Beutnagel et al., 1999) or even smaller units (Donovan and Woodland, 1995), though there is not a scale of which one sounds better. In (Kishore and Black, 2003) it was shown that syllables sound better than half-phones that in turn sound better than diphones and phones. However, this heavily relies on the database properties and the coverage it offers for a specific domain. Things slightly differ when moving from the spectral to the tonal domain, as the main point of F_0 unit-selection is to leave the original intonation "as-is" as possible. Long F_0 selection units should be preferred as they preserve the suprasegmental structure resulting in more natural intonation in cases of matching units. On the other hand they are hard to align with mis-sized target units, so smaller units are usually chosen (Meron, 2001; Raux and Black, 2003).

A different approach has been followed in (Xydas and Kouroupetroglou, 2004) where the F_0 curves are captured not in their whole range above each unit but only in the most prominent portions. The copy of original pitch samples in the database was combined with an onset/offset structuring of utterances, allowing higher F_0 definition in the accented words that are close to the utterance boundaries. The proposed pattern matching featured the alignment of stored pitch targets over phrases of different length, thus there was no need for storing samples of all possible syllable and pattern combinations providing high compression to the database. Though it lacked the naturalness in cases of long sentences, where an interpolation was applied, it proved successful for short messaging applications.

1.3. Modeling of expressive speech

Expressive speech synthesis is traditionally treated as an autonomous prosodic dimension that represents emotions like anger, joy, fear, happiness etc. (Schroeder, 2001; Eide et al., 2003), optionally followed by voice quality modifications (Gobl and Chasaide, 2003). Most efforts to model expressions propose the built of one database per expression. Emphasis can be considered as a primitive expression. The most common rule-based approach to emphasis realization is to insert pauses before and after the emphasized words, extend the duration and intensify the F_0. This sounds adequate for simple cases; however, naturalness is a step beyond adequateness (Black, 2003). Several works have dealt with modeling emotions in the intonational only dimension. The IPO model ('t Hart et al., 1990) has been proved to be able for identifying emotions in that manner (Mozziconacci, 2000). In (Pitrelli and Eide, 2003), a corpus-based approach was followed for modeling contrastive emphasis based on ToBI sequences. The experiments showed that this method was not strong at producing unambiguous percept of emphasis. One possible reason might be the fact that ToBI marks alone do not distinguish between emphasis or not. There are also some arguments (Mozziconacci and Hermes, 1999) that no specific intonational pattern is closely related with any emotion.

The unit-selection approach followed in (Raux and Black, 2003) has also been applied in emphasis

modeling, where subjective evaluation was done by comparing the proposed model against the standard rule-based one of Festival (Black et al., 1998).

1.4. Our work

Some prosodic phenomena, such as emphasis, which is examined in this work, cannot be applied to smaller chunks than a whole intonational unit like the breath group (Lieberman, 1967), the intonational phrase (Pierrehumbert, 1980) and the intonational group (Malfrere et al., 1998). Clark (2003) has also followed a looser definition of tone group considering it to be a "sequence of tones ending in some kind of boundary and nothing more". Our approach is closer to Ladd's one (Ladd, 1986), who defined a tone group as the smallest phonological unit that contains only one nuclear accent. In this work, we define a Tone-Group as the phonological unit that contains either a single pitch accent event or a single pitch accent followed by phrase or boundary tone event. To balance between keeping the database size to a minimum and making the unit size as big as required to capture at least one intonational group, this work proposes an F_0 model that is based on a perceptually acceptable partitioning of F_0 contours over the inventory's units. This minimizes both the database size and its creation effort and ensures sufficient coverage, as each pattern has now a more abstract structuring that enables it to achieve more than a single match, even with morphologically heterogeneous targets and thus it differs from the related approaches. We further apply this method to emphasis modeling and we are experimenting with the expressive effects that the chosen intonational unit size delivers.

In Section 2 we present the Tone-Group model and its selection algorithm. Section 3 describes the creation and the properties of an experimental database built from an expressive corpus. In Section 4 we perform a set of objective evaluations to measure model's performance in re-synthesis conditions, followed by a series of listening tests to compare the proposed model against a well-established linear regression one and to subjectively evaluate listeners' perception on distinct levels of emphasis (Section 5).

2. The Tone-Group model

The notion of a nuclear tone associated with a sequence of semantically consecutive words is the basis of several intonational theories like Ladd and Pierrehumbert's. Phrases that are set off by audible prosodic breaks are further broken down to units featuring single pitch accents. Based on that, a Tone-Group (TG) roughly extends to the boundaries of a clitic group (Nespor and Vogel, 1986). Namely, a TG is defined here as a sequence of consecutive words with word junctures of non-break type in between, where no more than one pitch accent occurs. Even though there is a certain confusion and disagreement regarding the terminology and definitions of the components of the prosodic hierarchy we can say that the TG – defined here as similar to the clitic group – also corresponds to what's been referred to in the literature as the minor phrase (Selkirk, 1978, 1986) and the prosodic word in its broader sense (cf. Selkirk's definition of a "super" prosodic word in (Selkirk, 1995)) (Fig. 1).

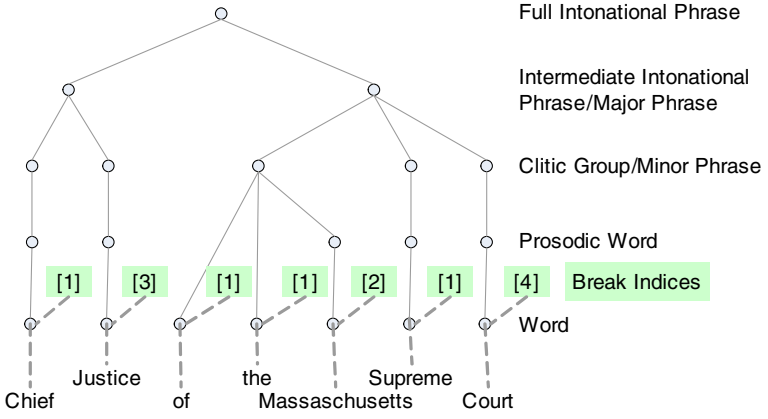

Fig. 1. Hierarchy of phonological scales for the utterance "Chief Justice of the Massachusetts Supreme Court." (Boston Radio News Corpus).

Moving to the tone domain, a TG is enriched – as mentioned above – by one accent tone and/or one endtone (either phrase accent or boundary tone). Other features (for example focus prominence information examined in this work) either weighted or non-weighted can be added in the feature vector in order to drive a finest selection. We will proceed to a closer inspection of the feature set in Section 2.2.

Prosodic databases that serve specific purposes differ from corpora, as the former are purpose-built collections of structured tokens that derive from the latter (Campbell, 2005). Taking into account the availability of corpora with sufficient linguistic labels and prosodic annotations, the proposed Tone-Group Selection model exploits existing information in order to build the model's database. For presentation and prototyping purposes we follow the utterance structure of the Heterogeneous Relation Graphs (HRG) representation (Taylor et al., 2001), though this is not compulsory. Accordingly, TGs are represented by the introduced ToneGroup relation that actually correspond to the clitic groups (Fig. 2). The ToneStructure tree relation allows the navigation from a TG down to the ToneItem one. Tone-Items carry the acoustic properties of the F_0 contour and are discussed in Section 2.3. Word and Syllable relations are also linked together by the SylStructure relation (not shown here) as happens in most common Festival voices. Fig. 2 shows this hierarchy of both the candidate units and the target ones. In case of a match, the ToneItem relation is copied from the candidate unit to the target, following the alignment process that is described in Section 2.6. According to the TG selection model, F_0 pitch targets, normally carried in the Target relation (not shown here), can be directly derived from the ToneItem relation. These targets describe the continuous pitch contour of the synthetic speech.

Selection of TGs is based on matching some target specification. Specifications are held as weighted feature vectors that include acoustic, linguistic and morphological information for all source and target TGs.

2.1. The 'haAb' pattern

To identify TGs, words in prosodic phrases are grouped and their syllables are further organized into patterns. In contrast to other related works, we do not model all the syllables within the range of a TG but only some key ones that correspond to major points, peaks or valleys on the pitch curve. For example, the pitch accent, phrase accents and boundary tones in ToBI. Since the construction of the patterns is based on the location and the type of accents and endtones, we suggest that the pro-

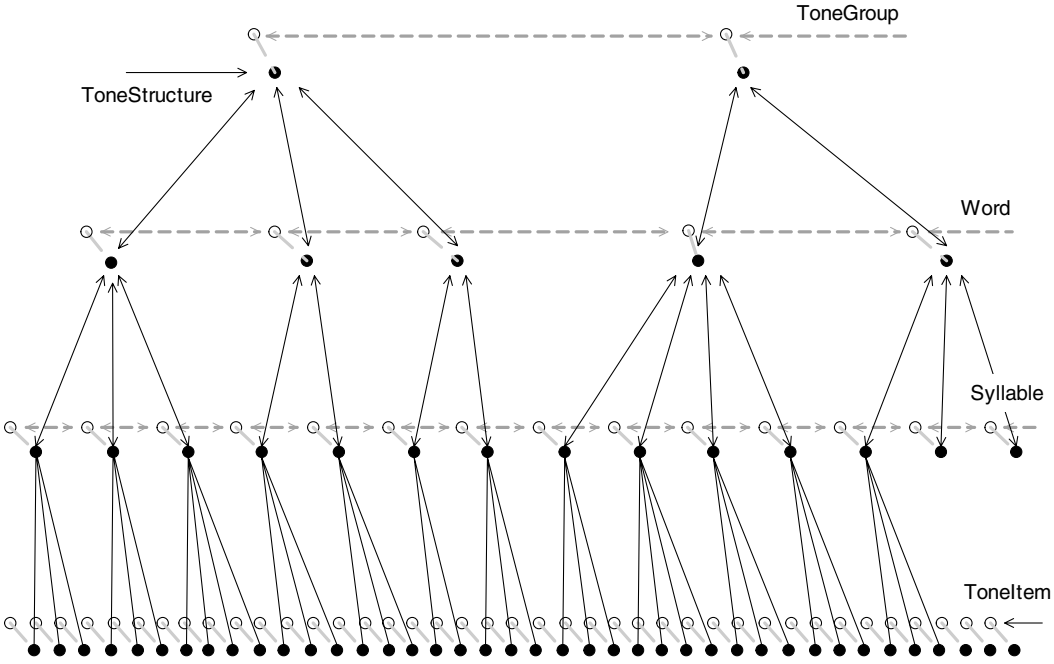

Fig. 2. The Tone-Group hierarchy as presented in the HRG.

sodic content of the curve is perceptually very close to the original when described by only these syllables, as it follows the tonal structure produced by the intonational model. Trying not to lose the F_0 definition, we are introducing an encoded pattern scheme that abstracts the morphology of patterns in order to keep the database size relatively small and also allow the efficient encoding of TGs in case of small-footprint applications.

Pattern construction can vary on how much detail we would like to represent. Our preliminary experiments showed that using up to four (4) syllables we can achieve high F_0 rendering. That was also validated by the high correlation between the modeled and the original pitch contours in the tests we performed (Section 4). Thus, we arrived to the following four (4) types of syllables:

- **h**: (or head) the very first syllable of the TG. It is the entry point to the TG.
- **a**: the first syllable of the word that carries the accent, if any (accent).
- **A**: the syllable that carries the accent, if any (accent).
- **b**: (or boundary) the last syllable of the TG. This is the exit point, where the F_0 curve ends. Normally, this syllable carries the endtone as well, if any.

In cases where any of the above overlaps during the labeling of a pattern, '**A**' and then '**a**' are preferred. These four syllables are called *haAb syllables*.

For example, the following sentence (taken from (Clark, 2003)) defines a TG that has three significant syllables according to our approach:

L* L-H%

I am a MILLIONAIRE?

h A b

and thus its identity is: {hAb;L*;L-H%}.

The basis of our method is that if a TG with a specific prosodic structure was spoken in one way, another TG with a similar structure should be spoken in a similar way in terms of speech synthesis, where prosody generation can hardly be unrestricted. Rooted in that, we suggest that the intonational pattern of the above TG can be successfully copied to any other TG with a matching identity.

Syllables that are not modeled and reside among the *haAb* ones are called *null syllables*. *Null* syllables are not sampled during the creation of the model and as we mentioned we suggest that there is no significant distortion to the perception of the contour. The consequences of this approach are discussed in the evaluation section, along with the effects of an inserted dynamic pitch interpolation within *null* syllables boundaries.

2.2. Feature set

The features that characterize a class are of great importance to either the accuracy of the classification in case of machine learning models or to the selection of the best matching unit in unit-selection. The most common feature set that is used in F_0 modeling includes linguistic (e.g. part-of-speech, syntactic chunk etc.), acoustic (e.g. ToBI accents, ToBI prosodic phrase break etc.) and other morphological (e.g. syllabic distance from major breaks etc.) properties. On the other hand, the prediction of the intonational events, which precedes the F_0 model in the TtS chain, incorporates almost the same set of features. For example, lexical stress is used to predict pitch accent and then lexical stress and pitch accent together are used to predict the pitch targets. This implies a redundancy on the selected features that sometimes biases the training procedure.

The feature set of a TG is divided in two sections. The first one, the primary features, forms its identity. These should make an exact match in order to allow a unit from the database to enter the candidate state. However, there are some exceptions to that like for example, if no unit is found to make a match or if the concatenation cost is above the desired thresholds. These cases are not discussed here, as preliminary evaluation of two 500-utterance sized databases (Xydas et al., 2005; Zervas et al., 2005) showed adequate coverage in all cases.

The primary features are:

- **pattern**: this is used to identify the syllabic morphology of the TG, as described before. Its values can be any sub string of "haAb".
- **accent**: declares the pitch accent of the TG. This can be either a categorical value like ToBI pitch accent classes (e.g. H^*, $H + L^*$) and rise/fall classes or continuous values like a Tilt event.
- **endtone**: declares the tone at the boundary of the TG. As previously, this can be either a categorical value like a ToBI boundary tone or phrase accent (e.g. L-L%, H-) or a continuous boundary value.

The above set is fundamental for "neutral" or "generic" speech synthesis. As we mention in Section 1 "Introduction", intonational patterns themselves cannot sufficiently represent expressive speech. For the purpose of emphasis modeling, we define one more feature to enrich the set that reflects the corresponding effect:

- **emphasis**: an indication of the focus prominence of the corresponding TG. Values can be either continuous or categorical (e.g. strong, moderate, none, reduced).

Features that belong to the second section, the secondary features, are weighted. This means that some are more important than others, and furthermore they are all less important than the primary ones. Secondary features provide a finer selection and can be ignored if required, as in small-footprint cases. Examples of secondary features include:

- **tg_in, tg_out**: in accordance to the commonly used syl_in/syl_out and word_in/word_out features, denoting the number of syllables or words since/until the previous/next major break, the tg_in and tg_out features indicate the number of TGs since or until the previous or the next major break. Both are numerical values.
- **p.accent**: the accent class of the previous TG.
- **p.endtone**: the endtone class of the previous TG.
- **n.accent**: the accent class of the next TG.
- **n.endtone**: the endtone class of the next TG.
- **num_of_syls**: the number of TG's syllables.
- **num_of_syls_in_last_word**: this refers to the word that usually carries the accent or the endtone and represents its number of syllables.
- **num_of_words**: the number of words in the TG.

Selecting and copying pitch contours between phonetically heterogeneous TGs can however affect the naturalness of the resulting speech. In (Aulanko, 1985; Whalen and Levitt, 1995; Monaghan, 1992; Vainio, 2001) has been pointed out the role of segmental prosody to the F_0 surface, such as the fundamental frequency difference between open and close vowels. In Greek, F_0 differences caused by the phonemic content do not follow the universal tendencies (Fourakis et al., 1999). In (Arvaniti et al., 1998) it was also shown that the alignment of the H target in the pre-nuclear $L^* + H$ accent is also affected by the segment type (stop, fricative, nasal) of the phoneme that precedes the post-accentual vowel. Such micro-prosodic variations is usually ignored in prosody generation, as either unit-selection speech synthesis does not distinct between tonal and spectral dimensions or other approaches cannot handle the overhead of the database size when trying to serve all the phenomena and combinations in speech.

To accommodate micro-prosodic variations, we allow the definition of the *micro-pattern* list in the secondary feature set that includes contextual information like the class (e.g. vowel, liquid), the identity and the quantity of segments. However, enabling this option causes a vast enlargement of the selection grid and this contradicts the small-footprint specification of the TGS model. Thus, we suggest that some additive methods can be followed in a more efficient way to correct micro-prosodic variations at runtime like the two-stage phonetic model that separates micro-prosodic events from prosodic phenomena and the residual component is added during synthesis (Monaghan, 1992), based on a *z*-score procedure (Bailly and Holm, 2005).

2.3. Tone items

To capture the human F_0 curve properties of a Tone-Group, several partitioning schemes of the curve can be followed (for example, every 10msec, as in (Keller and Keller, 2003)) and storage of the pitch values to apply them during synthesis. However, such an approach has the disadvantage that during synthesis time these points will need to be aligned over a different amount of segments (e.g. over a TG with more syllables than the stored one). Therefore, in order to capture all the possible combinations of segments, one will end up with a huge inventory. The terminus seems to be the segmentation of pitch contours and their abstraction so that combinations of small segments within an inventory can render back complete contours. Abstraction helps to generalize over non-significant portions of F_0 curves, like those we defined as *null* syllables and thus to compress the inventory.

A flexible, in terms of segmenting F_0, approach was followed initially in CHATR (Black and Hunt, 1996) and then in Festival (Black et al., 1998). Three Linear Regression (LR) models are constructed, one for pitch prediction at the beginning of syllables, one at middle of the vowel and one at the end of syllables (we will call this "the *sme* strategy" – *s*tart, *m*id, *e*nd). We keep this LR scheme as a reference because it is being widely used by speech researchers

and also we have experimented a lot in the past with it using our Greek corpora (Xydas et al., 2005). Moreover, there are legacy implementations that we can compare with in the evaluation section. The major drawback of this *sme* strategy is that a lot of intonational events occur away from the middle section of a vowel and the three single points cannot accurately render all accent types. For example, in $L^* + H$ pitch accent, the H tone has a late alignment in the accented vowel and the L^* a very early one, if not late in the previous syllable, whereas in $L + H^*$ the peak appears after the middle of the vowel (Arvaniti and Baltazani, 2005). This specification of the LR approach cannot be easily treated; if the mid-point was floating it wouldn't be possible to build the middle LR model, as there wouldn't be any positioning clue to place the target.

A more detailed perceptual encoding of pitch targets is described in the IPO intonation model ('t Hart et al., 1990). The original pitch contour is stylized from peaks to valleys and any variation that is subjectively judged not to be relevant to perception is filtered out. This procedure results to a set of straight lines that produce a perceptually equal copy of the original intonation when re-synthesized. The perceptual tests definitely set an overhead during the creation of the stylized contours. On the other hand, prosody does not always sound as good as statistics show, thus perceptual tests provide more confident results. To apply during synthesis, language-specific grammars are developed that represent standard pitch movements (for example for Dutch, English and German). These movements are characterized by their timing in the syllable, their spread over one or several syllables and their size relative to a top line. As has pointed in (Sproat, 1998), the possibility of the grammar to generate contours that are perceptually unacceptable is the most problematic side of the IPO approach. In contrast to IPO, we do not perceptually stylize the F_0 contour, but we just sample it "as-is". Instead of a grammar we use the selection algorithm to match contours that concatenate smoother. We only use perceptual criteria to distinguish between *haAb* and *null* syllables. However, this does not require any listening tests; the intonational analysis provides us with clues of where the accents and the endtones lie in utterances. Furthermore, we propose the modeling of specific syllables within a prosodic word in order to allow the reduction of the inventory's size.

To control the positioning and alignment of targets and cope with the aforementioned problems, we introduce the Tone-Item (TI), a dynamic list of parameters, which carry the acoustic properties of *haAb* patterns in the tonal dimension. TIs encode samples of pitch targets at any location. Though they are linked to the *haAb* syllables, the scope of their lists can be lengthened, if required, to describe specific intonational phenomena; for example, when the peak is positioned outside the accented 'A' syllable. As has pointed out in (Arvaniti et al., 1998), the canonical alignment of bitonal pre-nuclear $L^* + H$ accent in Greek can occur outside the accented syllables under certain circumstances and its alignment and positioning depend on the number of unaccented syllables between two accents. More specifically, a 10 ms mean interval has been measured from the onset of the first post-accentual vowel to the H target in case of words with lexical stress in the antepenultimate, as well as a 5 ms mean interval before the onset of the accented syllable where the F_0 minimum occurs (which, however, can be set in a reliable early point inside the accented syllable (Arvaniti and Baltazani, 2005)). In the former case, if 'A' is followed by a *null* syllable then this peak will be missed without the TI's list lengthening (actually, this can occur only in words with their lexical stress in the antepenultimate, otherwise a 'b' syllable will follow the 'A' and will capture the H peak).

2.3.1. TI features

Each pitch target is a node in the Tone-Item dynamic list. The following feature set accommodates each node:

- **syl_idx**: the syllabic index shows to which syllable the current target belongs to since the beginning of the linked *haAb* syllable. A positive value expands the list to the right side of the linked syllable whereas a negative value expands it to the left.
- **pos**: this feature shows the position of the target in the time axis. The default strategy provides a percentage value, relative to the duration of the voiced segments of the linked syllable indicated by the **syl_idx**.
- **f0**: the actual F_0 value for that **pos** in that **syl_idx**.

The F_0 value and positioning information provide a more appropriate placement of the targets during synthesis in contrast to the "sme" strategy. The syllabic index provides the list of TI the ability to be widened over one, two or more syllables. The

size and the density of targets of each list are arbitrary and depend on the curve shape and the required definition. Fig. 3 illustrates how the TIs encode a portion of an F_0 curve as well as the role of the syllabic index. Positioning of $TI_{i,j}$ is expressed in percentage of the duration of the voiced part of the syllable they belong to (based on the syl_idx).

2.3.2. The F_0 scanner

To capture f0 and pos values of TI lists, several sampling strategies can be applied during the creation of the TG model-specific database. The F_0 scanner is a process that samples a stylized copy of the contour above a specific unit. This can be done following three strategies:

1. **Threshold strategy**: samples are progressively selected so that the re-rendering of the contour should not exceed specific correlation, RMSE or other metrics thresholds compared to the original.
2. **Fixed strategy**: the contour is sampled on constant defined points. For example, in 0%, 20%,

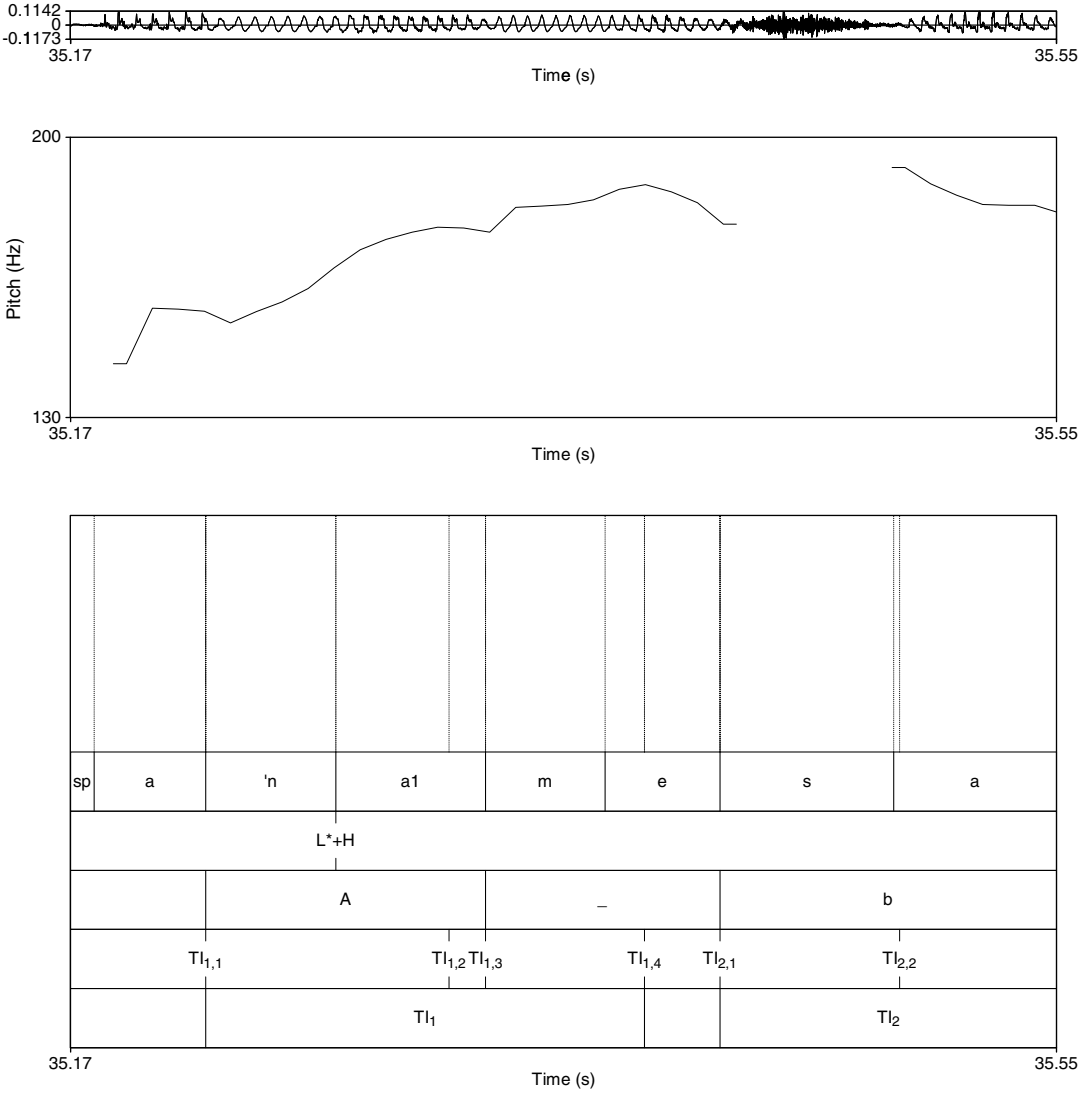

Fig. 3. The Tone-Items of the word /a'namesa/ (in between) that features an $L^* + H$ tone. Tiers (from top to bottom): segment, accent, *haAb* pattern syllables, Tone-Item list, and Tone-Item. As we see, the $L^* + H$ accent peaks in the first post-accentual vowel, i.e. 'e'. Thus, the TI list has been extended to the right, introducing $TI_{1,4}$ that otherwise would not exist as it links to a *null* syllable. $TI_{1,1} - TI_{1,3}$ have syllabic index of '1' whereas $TI_{1,4}$ has a '2' (not shown in the figure).

50% and 80% of the duration of voiced segments or in more descriptive locations like the "middle of the vowel".

3. **3-point strategy**: three pitch targets are selected from: (1) the beginning of the syllable, (2) the point where the F_0 peak (or valley) occurs and (3) the end of the syllable.

Fig. 4 shows an example of each strategy against the original contour.

Fig. 5 demonstrates an utterance divided in three TGs along with their corresponding TIs. On the bottom, there are the patterns of each TG aligned with the syllable relation: "haAb", "hAb" and "aA" respectively.

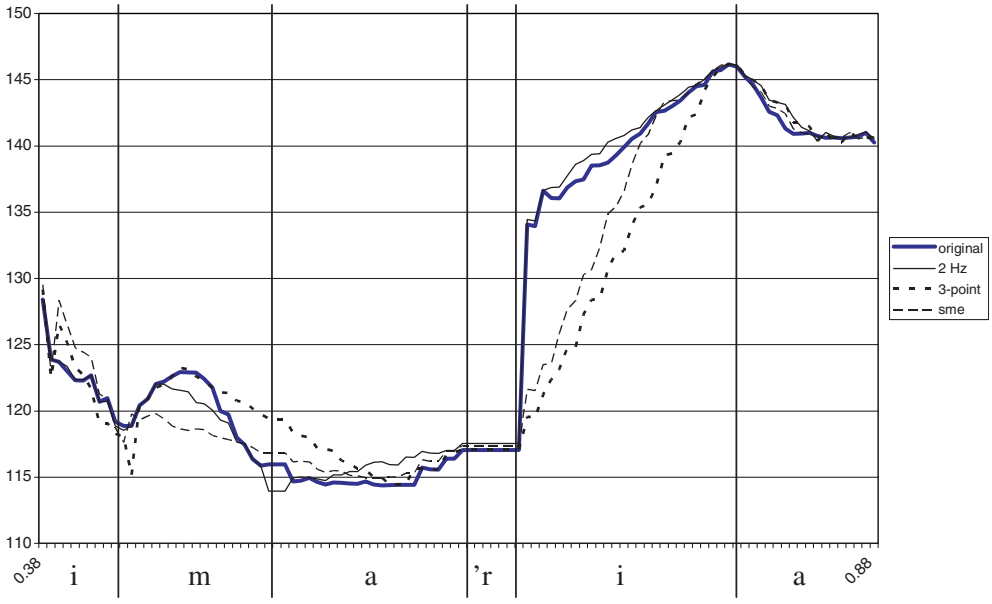

Fig. 4. Examples of re-synthesized pitch contours of the phrase /i ma'ria/ (Mary) based on the three different sampling strategies. The "2 Hz" is a threshold strategy where targets in the re-synthesized curve should not differ more than 2 Hz to the original. "sme" is the fixed strategy of *start*, *mid-vowel*, *end* of syllable that is followed in common linear regression approaches in Festival.

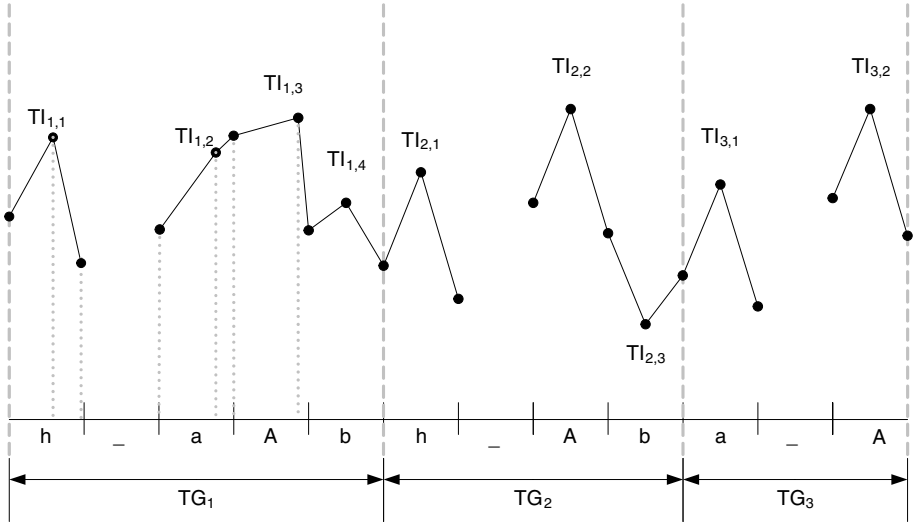

Fig. 5. Tone-Groups, *haAb* patterns and Tone-Items. The horizontal grid represents the syllables, whereas the dotted lines show the positioning details of each target. In TG_3 the 'h' and the 'a' syllable overlap, so it is marked as 'a'. The '_' represent the *null* syllables, i.e. those syllables that are not sampled.

2.4. Selection algorithm

Prior to the selection, we assume that a front-end system is able to parse utterances and provide the target TGs along with their feature vectors. This requires the prediction of the prosodic structure in terms of ToBI marks or other intonational notation. After that, a grid of candidate TG units is constructed. These are gathered by matching the primary feature vectors of units in the database with the target ones. Fig. 6 presents a grid.

In order for the grid not to grow through redundant TGs with similar acoustic properties that would consequently increase the required time for the selection, a pruning stage is applied after the construction of the grid. The features that are used as acoustic metrics to calculate the mean of a class and the distance of each candidate from the mean are:

- **start_f$_0$**: The F_0 value at the beginning of the TI.
- **max_f$_0$**: The maximum F_0 value in the TI.
- **end_f$_0$**: The F_0 value at the end of the TI.
- **mean_f$_0$**: The mean F_0 value in the TI.
- **stddev_f$_0$**: The standard deviation of F_0 values in the TI.

Candidates that their corresponding values are less than a deviation threshold are rejected, taking into account that a similar one already exists.

2.5. The cost function

An important factor that affects the quality of the resultant F_0 contour is the function that calculates the concatenation cost between two candidate TG units and thus the total distortion of the curve caused by discontinuities. This function determines the path in the grid that sounds smoother having a minimum of distortions in the contour. Since our main objective is to preserve the original values of the F_0 targets (though we can perform F_0 stylization to fix mismatches), we first use the F_0 difference measured in cents between the two adjacent Tone-Items of the TGs to be connected. Furthermore, the F_0 slope difference is used as a binary criterion to accept or reject a connection, as steep slopes in connections usually result in trembling voice (Fig. 7).

We have introduced a multi-pass search algorithm in the grid of candidate units. In each pass we increase the allowable concatenation threshold above which a connection is rejected. Therefore, in the first pass we look for an almost perfect curve with no discontinuities or steep slopes, in the second we allow small discontinuities to occur (which however can be stylized) and so on. Fig. 8 shows two TG candidate paths of a hypothetical phrase and their costs.

In that example, though the first path has a lower absolute cost, the second one is preferred as it features smaller standard deviation from an acoustically acceptable (according to the chosen threshold) mean difference. This way we achieve to select a path that provides an acoustical smooth connection from the one end to the other, by avoiding having steep discontinuities (like connection cost 8 in the above example). This multi-pass approach is time consuming, but it performs real-time in common embedded systems (e.g. 133 MHz ARM processor)

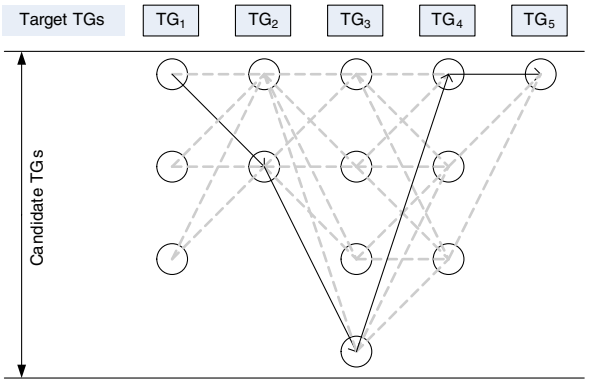

Fig. 6. The grid of candidate (circles) Tone-Group units. Each candidate matches the feature vector of the target unit on the top.

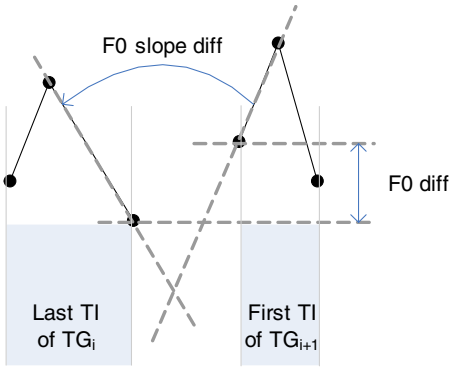

Fig. 7. Slope difference between two conjunctive Tone-Group units. As can be shown, the bigger the difference the acoustically smoother the connection. Black circles are independent pitch targets of the corresponding Tone-Items (TI).

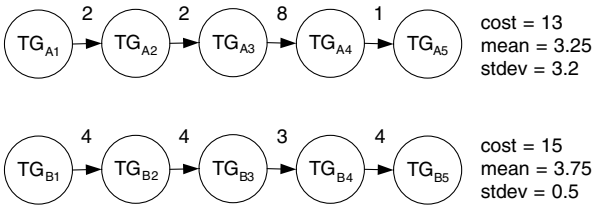

Fig. 8. Example of two candidate tone group paths.

due to the small primary feature vector. Also, higher pruning facilitates faster passes.

A symbolic code of the function that calculates the concatenation cost of a path is presented in Fig. 9. In each pass, the slope and the difference thresholds (SLOPE_THR and DIFF_THR) are increased according to the parameters of the model.

2.6. Alignment and positioning

One of the purposes of the TG patterns is to allow phrases to exchange pitch information between intonationally common syllables. Thus, after the calculation of the TG path, the next step is to align the source Tone-Items to the target TGs. This process is described in Fig. 10. This figure also depicts the abstracted property of a *haAb* pattern that allows it to match to morphologically different patterns.

The list of pitch targets in the source TI is copied upon the corresponding syllable of the target. This process actually generates the Target relation of HRG. For each individual pitch target, alignment is based on the positioning information carried by the corresponding node in the TI's dynamic list.

```
if slope_difference(TG_i, TG_i-1) < SLOPE_THR
        connection_cost = 1200*log2(first_tone_item(TG_i)/last_tone_item(TG_i-1))
        ACCEPT_PATH
else
        REJECT_PATH

if (ACCEPT_PATH) AND (connection_cost < DIFF_THR)
        path_cost = path_cost + connection_cost
else
       REJECT_PATH
```

Fig. 9. Path concatenation cost. The second line calculates the cents.

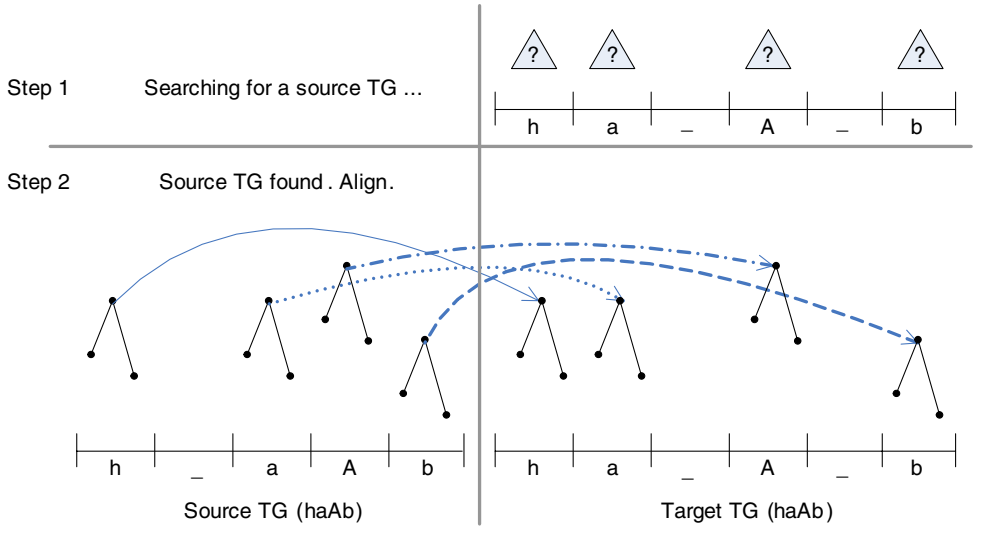

Fig. 10. Aligning the TIs from the TG in the database (left) to the target's one. Step 1 concerns the selection from the grid of candidates while the target TG (on the right) waits for four TIs to match to its *haAb* pattern. Step 2 presents the alignment.

Pitch is always positioned in sync to the percentage portion of the selected candidate, defined by the relative ratio of the target syllable's duration. The *micro-pattern* list offers clues to apply additive micro-prosodic corrections in this stage; however small-footprint prototyping does not include this due to the repercussions in the dataset size.

The syllables that have been characterized as intonationally non-important (i.e. *null* syllables) gain an interpolation between the regional F_0 values. According to the TG definition, no F_0 peak or valley should have appeared in those portions during the database creation, so this assumption does not eliminate F_0 perceptual description. This however was set as an issue during the listening tests we performed.

3. Database description

Prosodic corpora need to be carefully designed and recorded in controlled conditions in order to efficiently serve the purposes they are built for. Linguistically motivated annotations (e.g. ToBI) that are usually performed by hand, semi- or full-automatic methods need to be revised by experienced linguists for their accuracy and consistency. We have used an existing corpus with naturally emphasized speech that features sufficient prosodic coverage. This corpus was developed within the M-PIRO project (Calder et al., 2005) and it consists of 482 utterances, 59% of which are accompanied by enriched linguistic information. The domain of the database concerns a guided tour in a virtual museum and each utterance is part of an exhibit's description. Each description has one to three alternatives based on whether the visitor has already visited a relative exhibit, whether the delivered information is new or old and what is the status of the visitor. For example, if the term "archaic period" is being mentioned for the first time in a description, then it appears with strong emphasis. If it has been mentioned before, and so it is not "new" information to the visitor, then emphasis is decreased, depending on a "mentioned counter".

A Natural Language Generator generated the texts. Descriptions were marked-up with linguistic information using the SOLE scheme (Hitzeman et al., 1999). SOLE documents were parsed by DEMOSTHeNES TtS (Xydas and Kouroupetroglou, 2001) and emphasis specification was extracted from the provided features. Texts were further exported in HTML format that included visual instructions concerning the emphasis and its level over the corresponding parts. A professional actor read the texts following the printed guidelines, under the orders of a linguist who also ensured the consistency between written and spoken versions. Segmentation was performed using the HTK tool and expert linguists added ToBI notation. The corpus was also crosschecked to verify that the spoken emphasis levels were the expected ones as stated in the printouts; if that was not the case the linguists adjusted the written ones accordingly (which was an easier procedure than re-recording the utterances).

The Tone-Group processor created 3.726 TGs. Four features were assigned to each TG: pattern, accent, endtone and emphasis. Tables 1–4 present the class distribution of each of those features.

To test our assumption that we do not lose significant length of the contour using the encoded *haAb* patterns, we derived that the average length of TG is

Table 1
Distribution of *haAb* patterns in the collected TGs

Class	Occurrences	Percentage
haAb	539	14.47
hAb	549	14.73
haA	329	8.83
Ab	796	21.36
A	159	4.27
aA	517	13.88
aAb	767	20.59
hA	70	1.88

Table 2
Distribution of accents in the collected TGs

Class	Occurrences	Percentage
$L + H^*$	963	25.85
$H^* + L$	601	16.13
H^*	404	10.84
L^*	266	7.14
$L^* + H$	1081	29.01
none	411	11.03

Table 3
Distribution of endtones in the collected TGs

Class	Occurrences	Percentage
H-	501	13.45
L-	55	1.48
L-L%	467	12.53
H-H%	8	0.21
L-H%	7	0.19
none	2688	72.14

Table 4
Distribution of emphasis' levels in TGs

Class	Occurrences	Percentage
2	600	16.10
1	321	8.61
0	2805	75.28

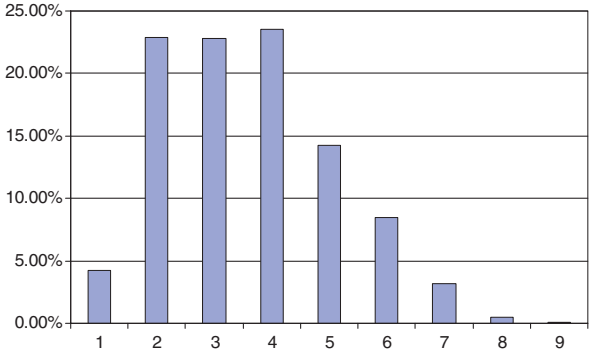

Fig. 11. The distribution of the syllabic size of TG patterns.

Table 5
Distribution of the TGs' number of *null* syllables

Num of Syls	Occurrences	Percentage
0	1703	45.71
1	1119	30.03
2	515	13.82
3	275	7.38
4	96	2.58
5	16	0.43
6	1	0.03
7	1	0.03

relatively short (Fig. 11) compared to the corresponding length of patterns as presented in (Malfrere et al., 1998). Also, Table 5 presents the distribution of the number of consecutive *null* syllables in TGs.

4. Objective evaluation

We first present an experimental analysis of our suggestion on the construction of the *haAb* and the *null* syllables. To do that, we performed a re-

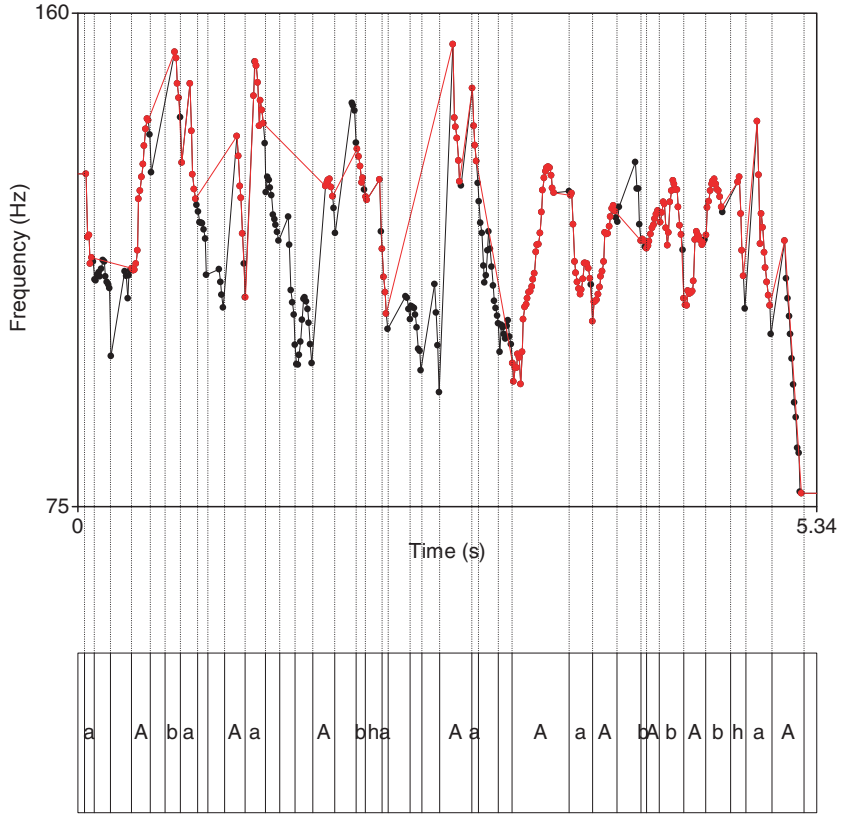

Fig. 12. Comparison of the original pitch contour (dark line) with the synthetic featuring straight-line interpolation over the *null* syllables.

synthesis of the original pitch curve based on the TG model, by replacing the original contours above the *null* syllables with straight lines. An example is shown in Fig. 12.

Following a full-scale sampling of *haAb* syllables, we calculated the correlation to be 0.94 and the root mean square error (RMSE) to be 15.23 between the original pitch and the one produced by the insertion of the straight lines above the *null* syllables. These values correspond to "what we miss" when we encode pitch contours following the *haAb* scheme.

Table 6
Correlation and RMSE of the synthetic pitch against the natural when (a) copying F_0 above *haAb* syllables using different sampling strategies and (b) introducing the "straight line" interpolation over the *null* syllables

Strategy	Correlation	RMSE
Full-scale (0 Hz threshold)	0.94	15.23
2 Hz threshold	0.92	16.61
3-point	0.89	24.28
Start, mid, end (sme)	0.89	24.27

Table 6 shows the correlation and RMSE in several sampling strategies.

As expected, the simplistic straight-line interpolation heavily affects the performance. On the other hand, since these parts do not represent any nuclear accent nor significant pitch movements, we assumed that a more linguistically aware rendering would enhanced the performance. The approach followed here is based on a scalable progression (rising or falling) of the pitch in between the boundaries of the *null* syllables, using data from the pitch baseline, the adjacent *haAb* syllables and the word morphology. Fig. 13 depicts the effect of this scalable pitch interpolation that replaces the straight lines above the *null* syllables with something more meaningful. This plain approach shows great improvement in the performance of the model (Table 7) and partially supports our assumption that we have managed to represent the pitch curve by modeling only the important syllables, whereas some reasonable pitch scales can successfully replace the rest portions. However, more investigation is required to

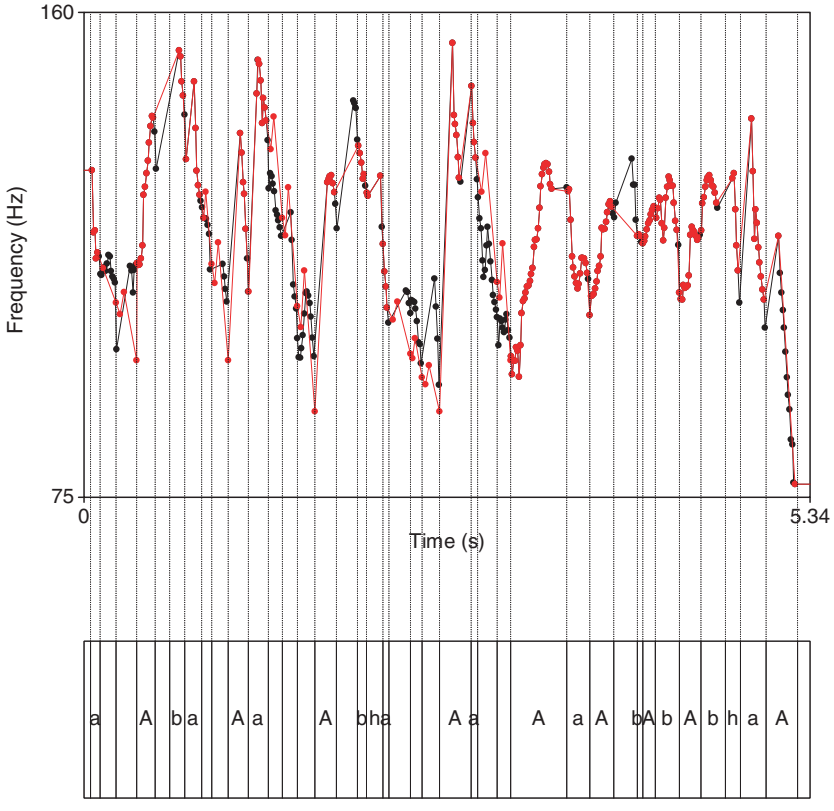

Fig. 13. Comparison of the original pitch contour (dark line) with the synthetic featuring scalable pitch interpolation over the *null* syllables that is based on pitch baseline and F_0 levels in the adjacent *haAb* syllables.

Table 7
Correlation and RMSE of the synthetic pitch against the natural when (a) copying F_0 above *haAb* syllables using different sampling strategies and (b) introducing the "scalable" interpolation over the *null* syllables

Strategy	Correlation	RMSE
Full-scale (0 Hz threshold)	0.96	9.20
2 Hz threshold	0.96	8.91
3-point	0.92	17.29
Start, mid, end (sme)	0.92	16.05

improve the performance of the run-time interpolation.

It is also interesting to note that following the *sme* strategy (followed in Festival's Linear Regression) for the whole range of utterances (*haAb* and *null* syllables) we measured 0.96 (correlation) and 9.77 (RMSE) respectively. That should be the ideal maximum performance of the Festival LR model when using our data.

5. Listening tests

Subjective evaluation of synthetic intonation heavily depends on the underlying segmental quality as well. We performed two experiments in order to take such effects into account. In the first one, we used a natural voice carrier where we applied the F_0 modifications that the TG selection model generated. In the second one, we used the MBROLA (Dutoit et al., 1996) diphone synthesizer and the Greek diphone database "gr2" (Xydas and Kouroupetroglou, 2001) featuring the same TG selection F_0 model.

A group of 31 listeners participated in these experiments. Six of them were speech experts while the rest were post-graduate students. In total, six of them were blind and had previous experience with the specific MBROLA voice (gr2) and TtS systems. Three of them did not know what TtS was all about.

Sentences were passed to the TtS system along with the required accent, endtone and level of emphasis using an XML-based mark-up scheme. Thus, no ToBI or other prediction related to intonational events took place, as this would add more complexity to the experiments.

We used DEMOSTHeNES TtS as a front-end system to parse the XML annotation, carry the remaining NLP processing and provide the prosody and the phonetic specification to the signal-processing component.

All the speech waves used for the tests can be found in http://www.di.uoa.gr/~gxydas/en/tone_group_selection.shtml

During the listening tests we did not exploit the *micro-pattern* lists (and thus we expected to have some implications on the produced speech). There were two reasons behind this decision: (1) the legacy Linear Regression model that we compared to did not had corresponding segmental features and (2) the implementation we did targeted a real model in small-footprint environments (source code is available for Flite) and such an extended feature set would exponentially increase the required inventory to perform the selection. In Section 6 we discuss the effect of micro-prosodic normalization that resulted during the segmental heterogeneous source to target pitch alignment.

5.1. Test setups

We selected 12 generic sentences that were out of the M-PIRO dataset and domain. The sentences were collected from (a) a newspaper (art and news), (b) spoken messages appearing in a mobile operator's customer services and (c) a literary book. Contrastively, the utterances from which the model was created belong to a museum domain with restricted grammar. We chose those tests as they fit better the "mobile domain" we targeted from small-footprint synthesis, though we do not define such domain restrictions. Since we had the prosodic specifications, our objective was to test the model in real conditions, and thus we chose sentences other than the already seen ones by the model. To confirm the domain differences of the test material, 79% of the content words of the tests were measured as lexically unknown to the M-PIRO dataset.

A professional speaker uttered these 12 sentences with no major pitch and intensity alterations in his voice, i.e. avoiding emphasizing or de-emphasizing any part of the utterance. These recordings were further segmented (automatically and hand-corrected) down to the phoneme level. In order to avoid perceptual annoyance usually caused by diphone-based synthesis or segmental discontinuities caused by bad selection in unit-selection synthesis and allow listeners to concentrate only on the intonation aspect of the experiments, we first chose to have perfect segmental quality by using a natural voice carrier for the delivery of the phonetic information, as a good unit-selection TtS would provide. To achieve this, the segmental information was passed to the

duration model in order to allow the alignment of the generated pitch targets with the original speech. A PSOLA (Moulines and Charpentier, 1990) algorithm was then used to modify the original speech signal according to the targets specification. This process is depicted in Fig. 14.

The second part of the experiments was carried out by using the MBROLA synthesizer and the gr2 database. In this case we tried to inspect whether user's perception of focus prominence remains the same when the segments have compressed and flat dynamics throughout the whole utterance, as happening in common diphone synthesis and small-footprint TtS. To aid the experimental purposes and the concentration of the listeners, we used the same original segmental information for the duration model as previously. This process is depicted in Fig. 15.

5.2. Against linear regression

The first listening test targeted the evaluation of the capability of the TG selection (TGS) approach to sufficiently represent the F_0 surface from a perceptual view. This was carried out to answer, "How natural does TG selection sound?" as we wanted to inspect how well the Tone-Items and the *haAb* pattern approach render the F_0 contour. For that reason, we set up a comparative listening test against a well-established Linear Regression (LR) model as has been presented in (Black and Hunt, 1996). According to this approach, three LR models are built for the beginning, vowel-mid and end of each syllable. The LR model, which we specifically used, had been also trained using data from the MPIRO corpus and had achieved good correlation and subjective comments (Xydas et al., 2005), while it has been widely used in legacy systems based on DEMOSTHeNES. Table 8 shows the correlation and the RMSE metrics performed by that model when predicting F_0 from ToBI marks and other commonly used linguistic features.

The subjects listened to the 12 synthesized stimuli that resulted from both the LR and the TGS model in random order. For each sentence they were asked to choose the most preferred one, LR or TGS. In both models, no emphasis information was provided in either the feature vector used for the

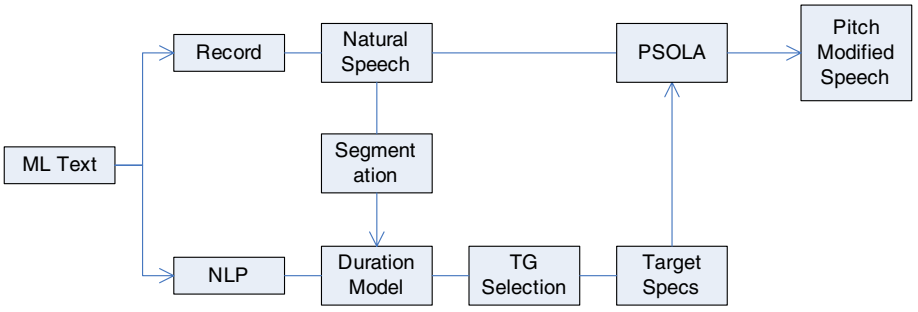

Fig. 14. The copy process of the synthetic selected pitch contour over a natural voice carrier. The mark-upped text on the left includes ToBI elements along with the text.

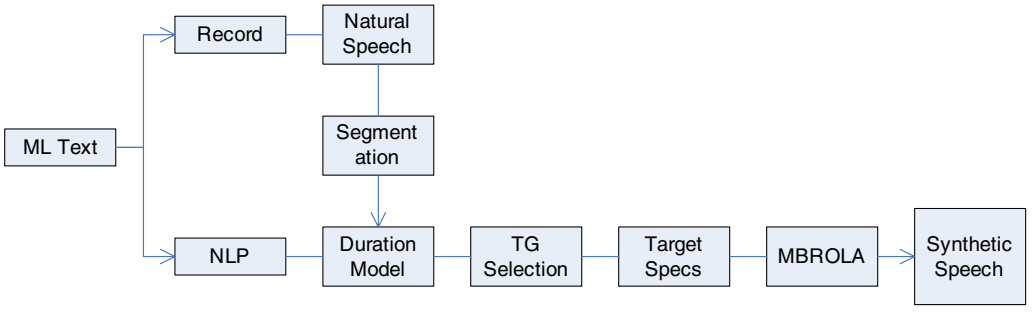

Fig. 15. The traditional TtS process with the natural speech segmentation.

Table 8
Correlation and RMSE of the linear regression model trained by the M-PIRO dataset (Xydas et al., 2005)

	Correlation	RMSE
Start	0.75	17.3
Mid	0.74	18.3
End	0.74	18.1

regression in the first case or for the selection in the latter. Thus, predicted values or selected units might incorporate emphasis, but not on purpose as source and target units were not compared based on emphasis.

Fig. 16 shows listeners' preferences in each stimulus when using a natural voice carrier. In total, the TGS model was preferred in 75% of the cases, while the LR in 25%.

Fig. 17 shows the listeners' preferences when using the diphone-based synthetic voice. In total, the TGS model was preferred in 74.4% of the cases against 25.6% of the LR.

Both cases clearly show that the TGS model sounded more natural to the users, though it models

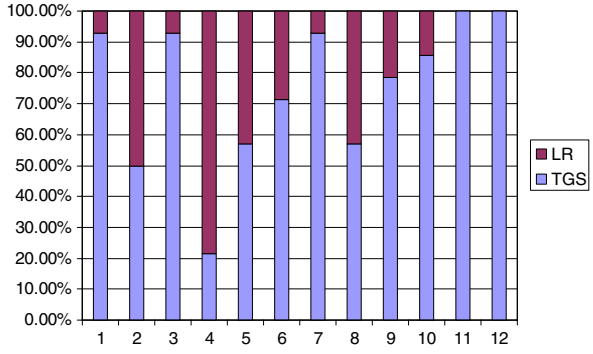

Fig. 16. Listeners' choices between LR and TGS, when using a natural voice carrier. Vertical axis shows the stimulus index.

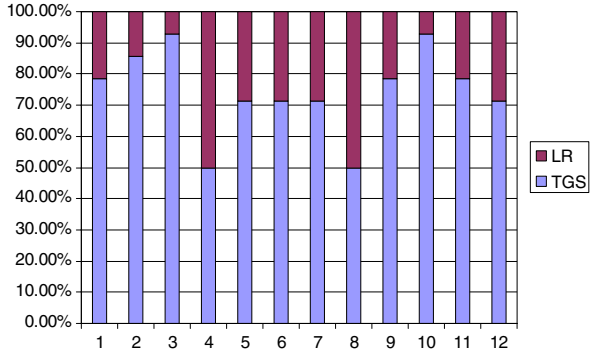

Fig. 17. Listeners' choices between LR and TGS, when using diphone concatenation. Vertical axis shows the stimulus index.

less syllables (only 4 at maximum per TG) than the LR model. This supports our assumption that we can encode only the intonationally most important syllables without losing in F_0 definition. Also, Figs. 12 and 13 show that no major difference was observed between the natural voice carrier and the diphone synthesis cases.

5.3. Emphasis rendition

The second experiment concerned the provision of emphasis to indicate focus in utterances. We used the same TG selection model. The difference from the previous setup was that we added the emphasis feature in the queries when looking for TGs to match with the targets. Emphasis was given in three levels, minor, major and none. Therefore, the TGS model performed the selection based on ToBI sequences as well as on the emphasis level. A natural voice carrier and the MBROLA diphone synthesizer were used as previously.

The stimuli were synthesized by defining the emphasis level over all the TG in the corresponding sentences. This means that the selected pitch levels were not in isolation but in relation to the whole sequence of *haAb* patterns that constitute each utterance. The purpose of the listening tests was to evaluate whether the resulting prompts could deliver the indented emphasis over each TG.

Each sentence featured at least one major focus and zero or more other prominent points of any level. At the beginning of the tests, two samples featuring the two different levels were presented in order for the listeners to get familiar with. After that, the subjects listened to each sample twice. For each sentence, they were given in advance two particular subphrases, *A* and *B*. They were asked to mark the level of emphasis they perceived at *A* and *B* as follows: 0 for null, 1 for minor and 2 for major. The null emphasis was included in order to ensure that listeners had the option to non-forcibly reject our intentions if they did not perceive any emphasis at all. Thus, they were instructed that "*there **might not** be any emphasis at some points, so put 0*" though this was never the case. This slightly differs from allowing them to vote 'null' when told that the possible values were only '1' and '2', which would seem to them as having to express a negative opinion.

The two different signal processing approaches were carried in two separate sessions, first the natural voice carrier with the modified pitch and then the diphone concatenated one. In each case we exam-

ined whether the subjects perceived any prominence and how close to the intended level their perception was. For that purposes we tested the null hypothesis that the listeners perceived the target emphasis in more than 1 level of difference. Thus, we calculated the probability p that the mean absolute difference M would be as different or more different from 1. We also made the assumption that our data are normally distributed.

Table 9 presents listeners' mean perceived emphasis (MPE) level, the mean absolute difference between the target emphasis and MPE, its standard deviation and the probability p mentioned above for each one of the utterances. Looking at the probabilities we can reject the null hypothesis in the majority of the cases ($p < 0.05$), though there are some (7 out of 24) exceptions as shown in the table. The reasons behind these exceptions might be (a) bad selection either because of the algorithm or because of the heterogeneity between the database and the testing domain, (b) the relative emphasis between points A and B that listeners perceived, (c) the semantics of the texts that might have imposed different points of emphasis and (d) the omission of the *micro-pattern* list from the selection feature set. The mean difference between the target emphasis and listeners' perception is 0.27.

Table 10 presents the same metrics when using the MBROLA voice. The mean difference between the target emphasis and listeners' perception is now 0.23. Also, the non-significant results now have been reduced to 5 (out of 24). It can be derived that

Table 9
Mean opinion score and stdev of the perceived level of emphasis for each stimulus (natural voice carrier)

Utterance	Point	Target emphasis	MPE	M	$\sigma(M)$	p
1	A	2	1.74	0.26	0.44	<0.05
	B	2	1.84	0.16	0.37	<0.05
2	A	2	1.80	0.19	0.40	<0.05
	B	1	0.67	0.35	0.49	0.09
3	A	1	0.83	0.29	0.46	0.06
	B	2	1.77	0.23	0.43	<0.05
4	A	2	1.70	0.32	0.54	0.10
	B	1	0.27	0.74	0.44	0.28
5	A	2	1.73	0.26	0.51	0.07
	B	2	1.73	0.26	0.44	<0.05
6	A	2	1.77	0.23	0.43	<0.05
	B	1	0.80	0.26	0.44	<0.05
7	A	1	1.17	0.23	0.43	<0.05
	B	2	1.80	0.19	0.40	<0.05
8	A	2	1.80	0.19	0.40	<0.05
	B	1	0.77	0.26	0.44	<0.05
9	A	2	1.90	0.10	0.30	<0.01
	B	1	0.87	0.19	0.40	<0.05
10	A	2	1.87	0.13	0.34	<0.01
	B	2	1.73	0.26	0.44	<0.05
11	A	2	1.70	0.29	0.46	<0.05
	B	1	0.53	0.45	0.51	0.14
12	A	1	0.67	0.32	0.48	0.08
	B	2	1.73	0.26	0.44	<0.05

Table 10
Mean opinion score and stdev of the perceived level of emphasis for each stimulus (diphone voice – mbrola, gr2)

Utterance	Point	Target emphasis	MPE	M	$\sigma(M)$	p
1	A	2	1.81	0.19	0.40	<0.05
	B	2	1.90	0.10	0.30	<0.01
2	A	2	1.90	0.10	0.30	<0.01
	B	1	0.80	0.23	0.43	<0.05
3	A	1	0.93	0.19	0.40	<0.05
	B	2	1.77	0.23	0.43	<0.05
4	A	2	1.73	0.29	0.46	<0.05
	B	1	0.60	0.42	0.50	0.12
5	A	2	1.80	0.19	0.40	<0.05
	B	2	1.53	0.45	0.51	0.14
6	A	2	1.87	0.13	0.34	<0.01
	B	1	0.87	0.19	0.40	<0.05
7	A	1	1.30	0.29	0.46	0.06
	B	2	1.83	0.16	0.37	<0.05
8	A	2	1.77	0.23	0.43	<0.05
	B	1	0.50	0.52	0.51	0.17
9	A	2	1.97	0.03	0.18	<0.01
	B	1	0.87	0.13	0.34	<0.05
10	A	2	1.93	0.06	0.25	<0.01
	B	2	1.77	0.23	0.43	<0.05
11	A	2	1.77	0.23	0.43	<0.05
	B	1	0.73	0.26	0.44	<0.05
12	A	1	0.77	0.23	0.43	<0.05
	B	2	1.67	0.32	0.48	0.08

listeners perceive 15% clearer the focus information in diphone synthesis (not diphone-selection), where the intensity is constant throughout the whole stimuli and the signal sounds like having flat dynamics.

6. Discussion

The alignment process we suggest in this paper takes into account the small-footprint specifications. A main concern is whether and how successfully can we stretch some source pitch points to the corresponding target syllables with mismatching durations and heterogeneous phonemic content. To inspect on the consequences of omitting the *micropattern* list (to minimize the inventory), we analyzed further the dataset of M-PIRO. Our goal was to measure differences in F_0 peaks when a particular accent of a specific phoneme is copied to the same accent over a different phoneme. For each *haAb* pattern of the collected TGs, we calculated the F_0 peak (or valley, depending on the type of accent) on the accented vowel. We further classified mean F_0 peaks based on the accent and the phoneme (Fig. 18).

Furthermore, Table 11 presents the standard deviation of the mean F_0 peak of each phoneme per pitch accent. From there, we can argue that by not accounting for micro-prosody we miss in average 8 Hz of the peak of each pitch accent. However, these findings should be compared with the actor's mean F_0 (= 152 Hz) and stdev (= 32 Hz) (Xydas et al., 2005), which attest a big variability in his voice.

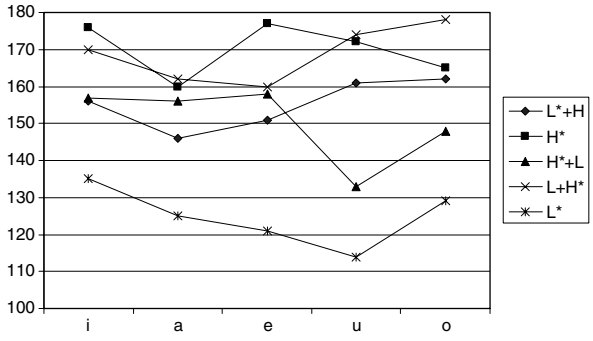

Fig. 18. Mean F_0 peak of the inventory's *haAb* patterns for each pitch accent.

Table 11
F_0 average peak and stdev across all the vowels in the inventory

Accent	Peak average (Hz)	Peak stdev (Hz)
$L^* + H$	155	6.8
H^*	170	7.3
$H^* + L$	150	10.5
$L + H^*$	169	7.7
L^*	125	7.9

7. Conclusions

Corpus-based F_0 modeling faces the problem of coverage in the database in order to provide the best of its intentions. This problem is even broader in cases of small-footprint conditions where the inventory is significantly reduced. In order to allow the provision of emphasis in speech synthesis that essentially increases the naturalness of speech, we proposed the Target-Group Selection F_0 model that encodes the intonational templates based on the intonational important syllables of each TG. The subjective experimentation that followed showed that this model is preferred against the well-established linear regression one. Furthermore, listeners were able to perceive emphasis where intended, however, the levels they distinguish differ, as emphasis perception is a subjective factor and people recognize it differently. The proposed approach achieved to deliver the two distinct levels and also to distinguish them from null cases.

The TG model achieves high coverage due to the encoding of the proposed *haAb* patterns. Only 8 pattern classes are provided (Table 1) that upon the application of a typical 482-sized utterance corpus gave a minimum of 70 instances in the lowest-frequent appearing class. The current prototype has no provision in cases of missing units yet, though no such case appeared when tested against a second unseen set of 500 utterances. On the other hand, as *haAb* patterns themselves offer 100% coverage, missing units can be handled by backing off to the closest ToBI label until a match is achieved.

In this work we inspected on a natural sounding F_0 model that can be fitted in embedded systems. The basis of the methodology described proved successful through prototyping and experimentation, and can form a framework for a combined pitch/duration/intensity selection based model.

Finally, an open-source prototype of the Tone-Group Selection model for the small-footprint Flite speech synthesizer (Black and Lenzo, 2001) is available from the project's web page[1] where the reader can also find the samples of the listening tests.

[1] http://www.di.uoa.gr/~gxydas/en/tone_group_selection.shtml.

Acknowledgements

The authors would like to thank George Frantzeskakis from the Hellenic Broadcasting Corporation (ERT) for offering his professional voice during the experimentation stage. The work described in this paper has been partially supported by the M-PIRO project of the IST Programme of the European Union under contract no IST-1999-10982.

Furthermore, we would also like to thank the reviewers for their constructive and useful comments.

References

d'Alessandro, C., Mertens, P., 1995. Automatic pitch contour stylization using a model of tonal perception. Comput. Speech Language 9, 257–288.

Arvaniti, A., Baltazani, M., 2005. Intonational analysis and prosodic annotation of Greek spoken corpora. In: Sun-Ah Jun (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, pp. 84–117.

Arvaniti, A., Ladd, D.R., Mennen, I., 1998. Stability of tonal alignment: the case of Greek prenuclear accents. J. Phonetics 26, 3–25.

Aulanko, R., 1985. Microprosodic features in speech: experiments on Finnish. In: Aaltonen, O., Hulkko, T. (Eds.), Fonetiikan Paivat Turku 1985, Publications of the Department of Finnish and General Linguistics of the University of Turku, pp. 33–54.

Bailly, G., Holm, B., 2005. SFC: A trainable prosodic model. Speech Comm. 46, 364–384.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Sydral, A., 1999. The AT&T Next-Gen TTS system. In: Proc. Joint Meeting of ASA, EAA and DAGA, Berling, Germany, pp. 18–24.

Black, A.W., 2003. Unit Selection and Emotional Speech. In: Proc. EUROSPEECH-2003, Geneva, Switzerland, pp. 1649–1652.

Black, A.W., Hunt, A., 1996. Generating F_0 contours from the ToBI labels using linear regression. In: Proc. ICSLP-96, Philadelphia, USA, Vol. 3, pp. 1385–1388.

Black, A.W., Lenzo, K.A., 2000a. Limited domain synthesis. In: Proc. ICSLP-2000, Beijing, China, Vol. 2, pp. 411–414.

Black, A.W., Lenzo, K.A., 2000b. Building voices in the Festival speech synthesis System. Available from: <http://festvox.org/bsv>.

Black, A.W., Lenzo, K.A., 2001. Flite: a small fast run-time synthesis engine. In: Proc. SSW4 – 4th ISCA Workshop on Speech Synthesis, pp. 204–207.

Black, A.W., Lenzo, K., 2003. Optimal utterance selection for unit selection speech synthesis databases. Internat. J. Speech Technol. 6 (4), 357–363.

Black, A.W., Taylor, P., Caley, R., 1998. The FESTIVAL speech synthesis system. Available from: <http://www.festvox.org>.

Bulyko, I., Ostendorf, M., 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In: Proc. ICASSP-2001, Vol. 2, pp. 781–784.

Calder, J., Melengoglou, A.C., Callaway, C., Not, E., Pianesi, F., Androutsopoulos, I., Spyropoulos, C., Xydas, G., Kouroupetroglou, G., Roussou, M., 2005. Multilingual personalised information objects. In: Stock, O., Zancanaro, M. (Eds.), Multimodal Intelligent Information Presentation, Text, Speech and Language Technology, Vol. 27. Springer, pp. 177–201.

Campbell, N., 1994. Prosody and the selection of units for concatenation synthesis. In: Proc. SSW2 – 2nd ESCA/IEEE Workshop on Speech Synthesis, NY, USA, pp. 61–64.

Campbell, N., 2005. Developments in corpus-based speech synthesis: approaching natural conversational speech. IEICE Trans. Inf. Syst. E88-D (3), 376–383.

Clark, R., 2003. Generating synthetic pitch contours using prosodic structure. Ph.D. dissertation, University of Edinburgh.

Conkie, A., Isard, I., 1994. Optimal coupling of diphones. In: Proc. SSW2 – 2nd ESCA/IEEE Workshop on Speech Synthesis, NY, USA, pp. 119–122.

Donovan, R., Woodland, P., 1995. Improvements in a HMM-based speech synthesizer. In: Proc. EUROSPEECH-95, Madrid, Spain, Vol. 1, pp. 573–576.

Dusterhoff, K., Black, A., 1997. Generating F_0 contours for speech synthesis using the tilt intonation theory. In: Botinis, A., Kouroupetroglou, G., Carayiannis, G. (Eds.), Intonation: Theory, Models and Applications. Proc. ESCA Workshop, Athens, pp. 107–110.

Dutoit, T., 1997. An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vreken, O., 1996. The MBROLA Project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In: Proc. ICSLP-96, Philadelphia, Vol. 3, pp. 1393–1396.

Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., Pitrelli, J., 2003. A corpus-based approach to expressive speech synthesis. In: Proc. SSW5 – 5th ISCA ITRW on Speech Synthesis, Pittsburgh, PA, USA, pp 79–84.

Fourakis, M., Botinis, A., Katsaiti, M., 1999. Acoustic characteristics of Greek vowels. Phonetica 56 (1–2), 28–43.

Gobl, C., Chasaide, A.N., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Comm. 40 (1–2), 189–212.

't Hart, J., Collier, R., Cohen, A., 1990. A Perceptual Study of Intonation – An Experimental-Phonetic Approach to Speech Melody. Cambridge University Press, Cambridge.

Hitzeman, J., Black, A.W., Mellish, C., Oberlander, J., Poesio, M., P. Taylor. 1999. An annotation scheme for concept-to-speech synthesis. In: Proc. 7th European Workshop on Natural Language Generation,, Toulouse, France, pp. 59–66.

Huang, X., Acero, A., Adcock, J., Hon, H., Goldsmith, J., Liu, J., Plumpe, M., 1996. Whistler: a trainable text-to-speech system. In: Proc. ICSLP-96, Philadelphia, PA, pp. 659–662.

Hunt, A., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP-96, Vol. 1, pp. 373–376.

Keller, E., Keller, B.Z., 2003. How much prosody can you learn from twenty utterances? Linguistik Online 17 (5), 57–79.

Kishore, S.P., Black, A.W., 2003. Unit size in unit selection speech synthesis. In: Proc. EUROSPEECH-2003, Geneva, Switzerland, pp. 1317–1320.

Ladd, D.R., 1986. Intonational phrasing: the case for recursive prosodic structure. Phonology 3, 311–340.

Lieberman, P., 1967. Intonation, Perception and Language. MIT Press, Cambridge, MA.

Malfrere, F., Dutoit, T., Mertens, P., 1998. Automatic prosody generation using supra-segmental unit selection. In: SSW3 – 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Blue Mountains, Australia, pp. 323–328.

Meron, J., 2001. Prosodic unit selection using an imitation speech database. In: Proc. SSW4 – 4th ISCA ITRW on Speech Synthesis, Perthshire, Scotland, 113.

Monaghan, A.I.C., 1992. Extracting microprosodic information from diphones – a simple way to model segmental effects on prosody for synthetic speech. In: ICSLP-1992, Banff, Canada, pp. 1159–1162.

Moulines, E., Charpentier, F., 1990. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Comm. 9 (5/6), 453–467.

Mozziconacci, S.J., 2000. The expression of emotion considered in the framework of an intonation model. In: Proc. ISCA/ITRW on Speech and Emotion, Belfast, Northern Ireland, pp. 45–52.

Mozziconacci, S., Hermes, D.J., 1999. Role of Intonation Patterns in Conveying Emotion in Speech. In: Proc. Internat. Conf. of Phonetic Sciences, pp. 2001–2004.

Nespor, M., Vogel, I., 1986. Prosodic Phonology. Kluwer Academic Publishers, Dordrecht.

Pierrehumbert, J.B., 1980. The Phonology and Phonetics of English Intonation. Ph.D. dissertation, MIT.

Pitrelli, J.F., Eide, E.M., 2003. Expressive speech synthesis using American English ToBI: questions and contrastive emphasis. In: Proc. IEEE ASRU-2003, pp. 694–699.

Quazza, S., Donetti, L., Moisa, L., Salza, P.L., 2001. ACTOR: A multilingual unit-selection speech synthesis system. In Proc. SSW4 – 4th ISCA ITRW on Speech Synthesis, Pertshire, Scotland, paper 209.

Raux, A., Black, A.W., 2003. A unit selection approach to F_0 modeling and its application to emphasis. In: Proc. IEEE ASRU-2003, pp. 700–705.

Schroeder, M., 2001. Emotional speech synthesis: a review. In: Proc. EUROSPEECH-2001, Aalborg, Denmark, Vol. 1, pp. 561–564.

Schweitzer, A., Braunschweiler, N., Klankert, T., Mobius, B., Sauberlich, B., 2003. Restricted unlimited domain synthesis, In Proc. EUROSPEECH-2003, Geneva, Switzerland, pp. 1321–1324.

Selkirk, E., 1978. On prosodic structure and its relation to syntactic structure. In: Fretheim, T. (Ed.), Nordic Prosody 2. TAPIR, Trodheim.

Selkirk, E., 1986. On derived domains in sentence phonology. In: Phonology Yearbook, Vol. 3, 371–405.

Selkirk, E., 1995. The prosodic structure of function words. University of Massachusetts Occasional Papers 18: Papers in Optimality Theory, pp. 439–469.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: a standard for labeling English prosody. In: Proc. ICSLP-92, pp. 867–870.

Sproat, R. (Ed.), 1998. Multilingual Text-to-Speech Synthesis – The Bell Labs Approach. Kluwer Academic Publishers, Dordrecht.

Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model. J. Acoust. Soc. Am. 107 (3), 1697–1714.

Taylor, P., Black, A.W., Caley, R., 2001. Heterogeneous Relation Graphs as a mechanism for representing linguistic information. Speech Comm. 33, 153–174.

Vainio, M., 2001. Artificial neural network based prosody models for Finnish Text-to-Speech synthesis. Ph.D. thesis, University of Helsinki, Department of Phonetics.

Whalen, D.H., Levitt, A.G., 1995. The universality of intrinsic F_0 of vowels. J. Phonetics 23, 349–366.

Wightman, C., Syrdal, A., Stemmer, G., Conkie, A., Beutnagel, M., 2000. Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative speech synthesis. In: Proc. ICSLP-2000, Vol. 2, pp. 71–74.

Xub, Y., Sun, X., 2002. Maximum speed of pitch change and how it may relate to speech. J. Acoust. Soc. Am. 111 (3), 1388–1413.

Xydas, G., Kouroupetroglou, G., 2001. The DEMOSTHeNES Speech Composer. In: Proc. SSW4 – 4th ISCA ITRW on Speech Synthesis, Perthshire, Scotland, paper 206, pp. 167–172.

Xydas, G., Kouroupetroglou, G., 2004. An intonation model for embedded devices based on natural F_0 samples. In: Proc. ICSLP-2004, Vol. 1, pp. 801–804.

Xydas, G., Spiliotopoulos, D., Kouroupetroglou, G., 2005. Modeling improved prosody generation from high-level linguistically annotated corpora. IEICE Trans. Inf. Syst. E88-D (3), 510–518.

Zervas, P., Fakotakis, N., Kokkinakis, G., 2005. Development of a prosodic database for Greek speech synthesis. In: Proc. SPECOM 2005 – 10th International Conference on Speech and Computer, Patras, Greece, Vol. 2, pp. 603–606.