Perceptual evaluation of text-to-speech implementation of enclitic stress in Greek

S.-E. Fotinea, A. Protopapas, D. Dimitriadis and G. Carayannis

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
evita@ilsp.gr

Abstract

This paper presents a perceptual evaluation of a text to speech (TTS) synthesizer in Greek with respect to acoustic registration of enclitic stress and related naturalness and intelligibility. Based on acoustical measurements and observations of naturally recorded utterances, the corresponding output of a commercially available formant-based speech synthesizer was altered and the results were subjected to perceptual evaluation. Pitch curve, intensity, and duration of the syllable bearing enclitic stress, were acoustically manipulated, while a phonetically identical phrase contrasting only in stress served as control stimulus. Ten listeners judged the perceived naturalness and preference (in pairs) and the stress pattern of each variant of a base phrase. It was found that intensity modification adversely affected perceived naturalness while increasing perceived stress prominence. Duration modification had no appreciable effect. Pitch curve modification tended to produce an improvement in perceived naturalness and preference but the results failed to achieve statistical significance. The results indicated that the current prosodic module of the speech synthesizer reflects a good balance between prominence of stress assignment, intelligibility, and naturalness.

Introduction

In Greek, only one of the last three syllables of a word may be stressed (Antepenultimate Rule). As a result, a very common and interesting phenomenon is the one of enclisis of stress, in certain cases of a clitic being attached to its preceding word. Special treatment is demanded when intonation words formulated by use of clitics trigger the phenomenon of stress enclisis, which involves two stressed syllables in one intonation word. As Holton, Mackridge & Philippaki-Warburton (1997) note, "The phonological word is a domain in which under certain conditions, more than one stress may appear: a basic and a derived one." With the attachment of the enclitic grammatical word, the Antepenultimate Rule may be violated. To recover, a second stress appears in the word. For example, /tok'alima/ (the cover) → /tok'alim'amu/ (my cover). Violation of the Antepenultimate Rule is recovered by the appearance of the second stress, which is corrective and thus stronger (Holton, Mackridge & Philippaki-Warburton, 1997).

Stress enclisis occurs in the following situations:

1. When a noun, adjective, adverb, or verb is stressed on the antepenultimate and is followed by a weak personal pronoun belonging to the same phrase, then a secondary stress must be placed on the last syllable of the first word. For instance, /x'aris'emuto/ (give it to me).
2. When a verb in the imperative is stressed on the penultimate and is followed by two weak pronouns, a secondary stress must be placed on the pronoun nearer to the verb. For instance, /r'ikset'uto/ (throw it to him).
3. If a gerund stressed on the antepenultimate is followed by one or two weak pronouns, a secondary stress will be placed on the last vowel of the gerund. For instance, /r'ixnond'asmu/ (throwing to me).

Earlier work on the pitch curve evolution (Fotinea, Vlahakis & Carayannis, 1997; Fotinea, 1999) has revealed that double stress intonation words at a position that is not sentence final and in an affirmative, neutral stress way of expression, can be piece-wise linearly approximated by an F0 pattern called Double-Stress Introductory (Db-I) pattern, which displays two points of pitch rise. The pitch curve remains almost flat until the first stressed syllable, then a pitch rise is observed, and in the next (unstressed) syllable a slight declination follows that allows for a new (more evident) increase of pitch at the next stressed syllable lasting to the end of the intonation word. This is noticed also in other researchers' results (Arvaniti, 1992).

In this work we investigated the interdependency of all three prosodic parameters, namely intensity, duration and pitch curve evolution, for the special case of stress enclisis, in the context of automatic speech synthesis. A formant Text-To-Speech synthesizer in Greek that has been developed at ILSP and is commercially available ("Ekfonitis") was used for the evaluation of the implementation of stress enclisis. In this system, intensity, duration, and pitch can be independently controlled. A prosodic module is also available that assigns stress to appropriately selected syllables according to rules by manipulating all three parameters in unison. The objective of this study was to evaluate the existing rules and to further explore the parametric space for the acoustic realization of enclitic stress. To this end, a set of short utterances were created, specifically contrasting stress among lexical items with and without attached clitics. As detailed below, listeners judged the stress pattern intelligibility and the naturalness of the synthesized speech, as a function of combinations of the three prosodic parameters. An important objective for this study was to employ a number of different assessment measures in order to separate (subjective) perceived preference from accurate perception of an intended stress pattern.

Method

Acoustical measurements on natural recordings

In order to study the acoustic marking of enclitic stress assignment, a set of sentences were constructed that contrasted regular stress, enclitic stress, and no stress, in phonetically balanced phrases. The basic phrase was /k'anod'as tu to to pr'aɣma/ ("doing the thing to him"), perfectly contrasting only in stress with /k'anodas t'uto to pr'aɣma/ ("doing this thing"). Additional variations were recorded and studied (/k'anodas to pr'aɣμα/ – "doing the thing"; /k'anod'as tu t'uto to pr'aɣma/ – "doing this thing to

him"; /k'anod'as tu to pr'aγma/ – "doing the thing to him"), the results of which are not reported here. Three initial-context conditions were designed, differing in the number of unstressed syllables that preceded the main phrase. This was made possible with the use of prefixes /kse/ ("un-") and /ksana/ ("re-") on the initial verb (/k'anodas/, meaning "doing"). This resulted in /ksek'anodas-/ and /ksanak'anodas-/ variations of each of the above sentences.

All sentences were recorded twice by each of four native speakers of Modern Greek with standard pronunciation, two male and two female, aged between 22 and 31, with normal hearing. The recordings were digitized and analyzed with the Praat program, a "research, publication, and productivity tool" written by P. Boersma. The sentences were marked on syllable boundaries and fundamental frequency and intensity contours were derived and compared. Examples of such recordings and analyses are shown in Figures 2 and 3 below.

Formant synthesis and analysis of the synthetic speech

The same sentences described above, in all variations and contexts, were synthesized by the Ekfonitis program automatically based on their text form, which was entered manually in MS Word. Each synthetic production was saved in an audio file and then marked and analyzed with the Praat program in the same way the natural utterances were. Figure 1 below shows one example of the analysis of such a synthetic utterance.

For the perceptual evaluation of enclitic stress assignment we selected the sentence bearing enclitic stress that was phonetically identical to a regularly stressed sentence (presented in the preceding section). All three initial-context variants were synthesized and analyzed. They were time aligned on the onset of the main phrase (the /k/ burst of the /ka/ syllable) so that manipulations on the following portion of the signal could be automated and controlled to be identical in all three variants.

Construction of enhanced stimuli

The synthetic speech stimuli were modified along the three prosodic dimensions of interest, namely pitch, duration, and intensity, with the focus of the modification concentrated around the syllable receiving enclitic stress. Figure 1 shows the original synthetic utterance /k'anod'as tu to to pr'aγma/ and Figures 2 and 3 show the same utterance as spoken naturally by male and a female speaker, with similar tempo and phrase intonation.

Note that the pitch of the synthetic utterance closely follows the Db-I model (with linear interpolation between regions, of course) and greatly resembles the pitch curve of this natural utterance. However, the natural utterance differs slightly in the temporal alignment of the pitch valley signaling an upcoming stressed syllable. Specifically, the synthetic curve shows a shallower pitch fall preceding the stressed syllable and a shallower rise following it. Therefore, the pitch modification chosen for evaluation was to approach this natural utterance more closely by altering the pitch values in the neighborhood of the syllable receiving enclitic stress, reducing pitch before it and increasing pitch after it. Figure 4 illustrates the pitch modification by superimposing the original and modified pitch curves.

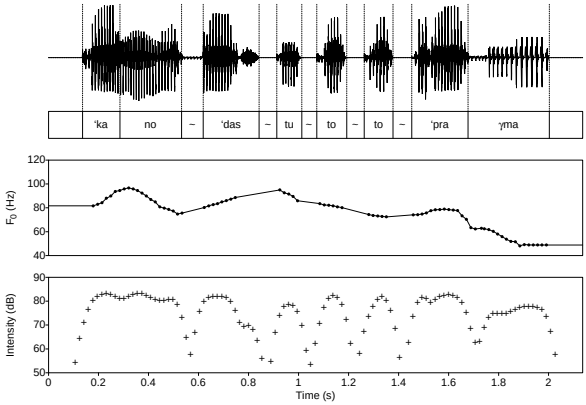

Figure 1: Original synthetic utterance /k'anod'as tu to to pr'aγma/ as produced by Ekfonitis.

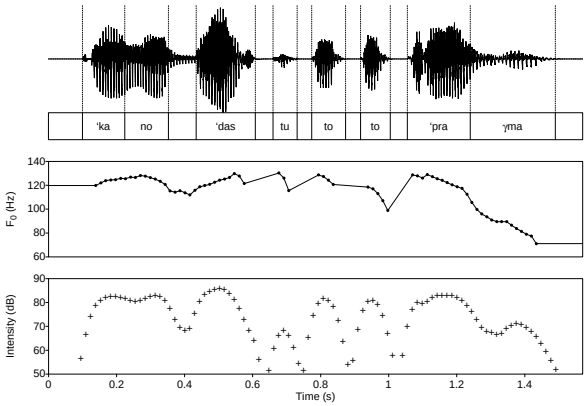

Figure 2: Natural utterance /k'anod'as tu to to pr'aγma/ spoken by male speaker M2.

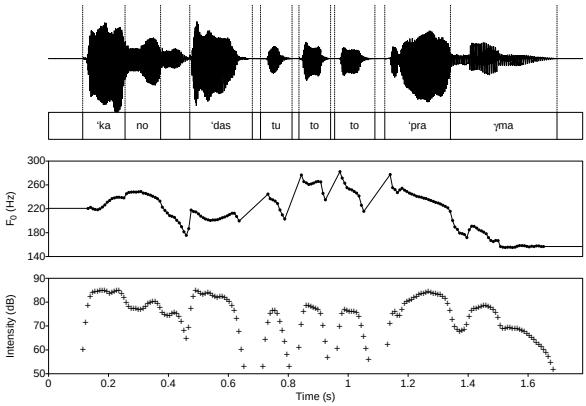

Figure 3: Natural utterance /k'anod'as tu to to pr'aγma/ spoken by female speaker F1.

Because duration is a known correlate of perceived stress in Greek (Botinis, 1989), the relevant modification consisted in an increase by 20% of the duration of the syllable receiving enclitic stress. Similarly, the intensity modification was performed by selectively amplifying by 3 dB the syllable receiving enclitic stress. Thus, overall, there were three different kinds of modifications, one for each dimension of interest, which resulted in eight stimuli

for each synthetic utterance: One unmodified, one for each kind of modification (pitch, duration, intensity) alone, one for each possible pair of dimension modifications, and one with all three types of modification. The corresponding resulting stimuli were given appropriate codes: 0, p, d, i, pd, pi, di, and pdi, respectively. The exact same modifications were performed for each of the three sentence length conditions, i.e., for zero, one, or two unstressed syllables preceding the first stressed syllable. Therefore, there were 24 stimuli based on the same set of original segmental and prosodic sequences with modifications along the three dimensions.

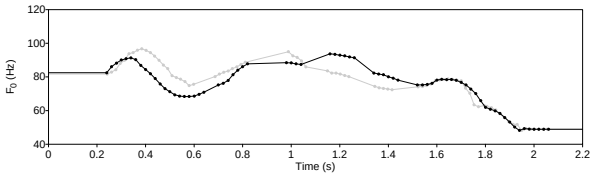

Figure 4: Original (gray) and modified (black) pitch curves for the synthetic phrase /k'anod'as tu to to pr'aγma/.

In addition to the 24 synthetic test phrases, a set of "distractor" stimuli were included in the data set and used in the perceptual assessment to verify that listeners paid attention to the stimuli and were not filling out the forms randomly. These additional stimuli were synthetic sentences that were segmentally identical to the test stimuli but differed prosodically and consequently in lexical segmentation and in meaning. Thus, for example, /k'anod'as tu to to pr'aγma/ contrasts with /k'anodas t'uto to pr'aγma/ with the meaning changing from "doing the thing to him" to "doing this thing." These stimuli were constructed for use in the stress assignment test.

Perceptual testing

There are several ways in which the modified synthetic stimuli might differ from the standard synthetic stimuli. Listeners may find them more natural or acoustically preferable; segmental and prosodic intelligibility may have improved or deteriorated; and perceived stress prominence may have also been affected. All of these considerations become relevant in an evaluation of synthetic speech and it cannot be known a priori which measurement might yield the most important information. Part of the present experiment was to employ a number of alternative measures of perceptual evaluation in order to study the diversity of evaluation considerations. Thus, following a brief familiarization session, the perceptual evaluation procedure was divided into two phases. Phase I included comparative perceptual assessment of naturalness and preference, that is, synthetic speech stimuli were presented in pairs and listeners judged which of each pair was more natural and which was preferable (without further specification of preference criteria). For this phase, only comparisons among pairs of stimuli differing by a single modification were selected, so that manipulations would be easier to assess individually. Phase II was designed to address specifically stress assignment and prominence, and each synthetic speech stimulus was presented individually. Listeners had to identify the syllables they perceived as stressed, by circling them on a syllabified printed version of the synthetic speech, and they had to indicate which of the stressed syllables sounded to them as bearing the heaviest (most prominent) stress. For this phase, all 30 synthetic phrases were used (standard, modified, and distractor stimuli). In accordance with the prosodic focus of the present study, effects on segmental intelligibility are thus not addressed in this testing situation.

Ten listeners, seven female and three male between the ages of 22 and 38, participated in the perceptual assessment task, individually or in groups. All were native Greek speakers and had no hearing impairments. Each received two rating sheets and was informed about the phrase-initial and phrase-final lexical content of the stimuli. That is, participants were not told what the critical syllables of the test phrases were or what intonation they should expect to perceive. In both phases, listeners were given the option to hear each stimulus or pair of stimuli a second time (but not more) before providing their responses. The stimuli were presented over a pair of loudspeakers at a normal, comfortable level. The order of stimulus presentation was randomized once prior to all ratings. For Phase I, the order of the two stimuli within each pair was also randomized at the beginning. The actual presentation of the audio stimuli was controlled by a custom program written in Matlab.

Results

All stimuli sounded very similar, and it was difficult to discern the modifications unless careful attention was paid to the stimuli when presented in immediate succession. Indeed, data collected for one of the ten listeners had to be excluded because there was no difference between trials. That participant claimed that she was unable to perceive any differences between stimuli at all. All other participants gave more varied responses, which were subsequently entered in Microsoft Excel for tabulation and statistical analysis. Table 1 shows the mean naturalness ratings for the 12 triplets of stimulus pairs, arranged by dimension of modification. The ratings are grouped over the three initial contexts (zero, one, or two unstressed syllables) and only mean values for the three contexts are shown. Because preference ratings were practically identical (only two data points out of 324 were different from naturalness ratings), they are not presented separately. Numbers close to zero indicate no preference while numbers with absolute value close to one indicate strong preference. Statistically, mean absolute ratings less than 0.41 are more than 5% likely to occur by chance (based on a binomial distribution with random equiprobable "preferences"). Therefore, in Table 1 there is one statistically significant preference and one marginal. The probability that one preference out of twelve total would come out at 0.56 if ratings were random is less than 1%, therefore we may accept that increasing duration over a synthetic utterance that has already been amplified decreases its naturalness. Although less reliable statistically, the decrease in naturalness for amplifying an utterance with already modified pitch and duration cannot be ignored. It should be also pointed out that the low reliability of the observed ratings might be justified by the great similarity among the test items and the relatively small number of listeners employed for the perceptual assessment test.

Condition 1	Condition 2	Modified parameter	Mean relative naturalness
0	d	d	0.19
0	I	i	-0.33
0	p	p	0.11
d	di	i	0.11
d	pd	p	-0.11
i	di	d	-0.56
i	pi	p	0.26
p	pd	d	-0.19
p	pi	i	-0.26
pi	pdi	d	0.33
pd	pdi	i	-0.41
di	pdi	p	0.33

Table 1. Mean relative naturalness (preference) ratings for the twelve comparisons. Each number represents the mean rating of nine subjects for the three initial-context stimuli in the modification conditions indicated in the two leftmost columns. Ratings range between –1 (indicating preference for condition 1) to +1 (indicating preference for condition 2).

Overall mean preference and naturalness judgments for the three kinds of modification indicate that intensity amplification is strongly disfavored (-0.22; p=0.016, two-tailed), pitch modification is somewhat favored (0.15; p=0.15), while duration lengthening has a negligible effect (0.06, p=0.63).

Regardless of preference or judged naturalness, the question of whether and how stress is perceived in each case must also be addressed. Table 2 shows the mean stress prominence ratings for each of the eight modification conditions (averaged over 9 subjects and over 3 initial contexts) and for the distractor sentence (listed as "t"), in which there was no enclitic stress but the following syllable was stressed instead. It is immediately apparent that perceived stress coincided with intended stress, since the mean stress rating of the syllable receiving enclitic stress for all eight variants is between 1 (stressed) and 2 (most prominent) whereas for the distractor sentence is nearly zero. In addition, it should be emphasized that the syllable bearing enclitic stress tends to sound most prominent (mean perceived stress > 1), in agreement with the theoretical prediction that enclitic stress is corrective and thus stronger.

Conversely, the following syllable is not perceived as stressed in the modification conditions (mean rating less than one) but only in the distractor stimuli, in which it was meant to be stressed (mean rating 1.22). The differences seen between the stress ratings in different modification conditions are not significant with the exception of the two extreme values. That is, the amplified syllable (i) sounds more stressed than the same syllable with all three modifications (pdi). It is, however, interesting to note that the intensity modification, which produced the least favorable ratings in the pair judgments in Phase I, has the most effect on perceived stress prominence.

Finally, the overall perception of stress assignment was measured by comparing the intended pattern of stressed syllables with the pattern indicated by the listeners in the response sheet. The observed differences between modification variants were negligible.

Condition (modified parameter)	Stress prominence for /das/	Stress prominence for /tu/
t	0,06	1,22
0	1,37	0,04
d	1,48	0,07
di	1,56	0,07
i	1,56	0,07
p	1,41	0,15
pd	1,37	0,07
pdi	1,30	0,22
pi	1,44	0,15

Table 2. Mean stress prominence ratings for the syllable receiving enclitic stress (das) and the following control syllable (tu). Each number represents the mean rating of nine subjects for the three initial-context stimuli in the modification condition indicated in the first column. Ratings range between 0 (indicating perceived unstressed) and 2 (indicating perceived as the most prominently stressed syllable); a value of 1 indicates perceived stressed but not as most prominent. (t: Distractor sentence.)

Conclusion

The implementation of the Db-I model by the Ekfonitis prosody module appears to be quite effective both in conveying the intended stress pattern and in sounding reasonably natural. Some of the modifications that were applied in this study were not perceptible whereas others affected perception of stress, often adversely. Specifically, intensity modification was disfavored despite increasing perceived prominence. Duration had no effect whereas pitch curve modification produced a slight preference which failed to reach statistical significance.

It is possible that more refined modifications to the prosodic module and larger groups of listeners could result in robust improvements in the naturalness of such complex prosodic manipulations. Further research is necessary to resolve this issue.

References

Arvaniti, A. (1992). Secondary stress: evidence from Modern Greek. In G. Docherty & D.R. Ladd (Eds.). Papers in Laboratory Phonology II: Gesture. Segment. Prosody (pp. 398–419). Cambridge University Press.

Botinis, A. (1989). Stress and Prosodic structure in Greek: A Phonological, Acoustic, Physiological and Perceptual Study. Lund, Sweden: Lund University Press.

Fotinea, S.-E., Vlahakis, M., & Carayannis, G. (1997). Modeling arbitrarily long sentence-spanning F0 contours by parametric concatenation of word-spanning patterns. In Proceedings of Eurospeech 97 (Vol.1. pp.315–318).

Fotinea, S.-E. (1999). Sentence-level Prosodic Modeling of the Greek language with Applications to Text-To-Speech synthesis. Doctoral thesis. National Technical University of Athens. Department of Electrical and Computer Engineering.

Holton, H., Mackridge, P., & Philippaki-Warburton, I. (1997). Greek: A comprehensive grammar of the Modern Language. London: Routledge.