Synthesising Hyperarticulation in Unit Selection TTS

Matthew P. Aylett

Scansoft Ltd
CSTR, University of Edinburgh
matthewa@inf.ed.ac.uk

Abstract

Within speech synthesis we often wish to give extra focus to words which carry important information, such as names, dates and amounts. In this paper we look carefully at cost functions that can be used to bias unit selection in favour of hyper-articulated speech in order to give this impression of focus. Hyper-articulated speech tends to be accented, emphatic and requires more articulatory effort. We apply two cost functions to try to force the selection of hyper-articulated speech. The first operates on the duration of units in the unit selection database, the second on the language redundancy (word trigram predictability) of the word containing the unit. We estimate their relative importance in selecting hyper-articulated speech in unit selection speech synthesis. A listening test was carried out where these cost functions were applied to one random content word in a haskins anomalous sentence. Listeners were asked to select the two clearest and most focused words from the sentence. The duration increasing cost function was significantly related to an increase in perceived prominence whereas low redundancy, and a combination of both approaches did not produce significant results. Thus, although a significant correlation exists between the average duration and redundancy of diphones and perceived prominence, such a correlation was not smoothly translated into error free method for altering such perceived prominence.

1. Introduction

Prosodic models in TTS systems have varied from rule based prescriptive models, based on an implicit or explicit knowledge base [1], to data driven models such as: CART decision trees trained from a speaker's data [2, 3], lazy learning approaches using tree matching e.g. [4], and unit selection based on a viterbi search [5]. However with the increasing use of very large unit selection databases it has become apparent that much prosodic structure can be synthesised 'for free' without any (or very minimal) direct reference to a global duration or f0 target. However, such systems, which have a very low cost attributed to global f0 and duration cost functions, are also difficult to manipulate. After all, if the model is mostly ignored, how do you affect a prescriptive change?

One requirement for TTS engines is to impose prominence or focus on part of the utterance in order to improve intelligibility of a particular word or phrase in order to convey new information. In this paper we look carefully at cost functions that can be used to bias selection in favour of hyper-articulated speech in order to give this impression of focus.

Recent work in phonetics has shown a marked correlation between prosodic prominence, duration and language redundancy [6, 7]. Language redundancy can be thought of as the chances of guessing a word given the context the word is in. A word trigram model, where a word is guessed given the previous two words approximates this sense of language redundancy.

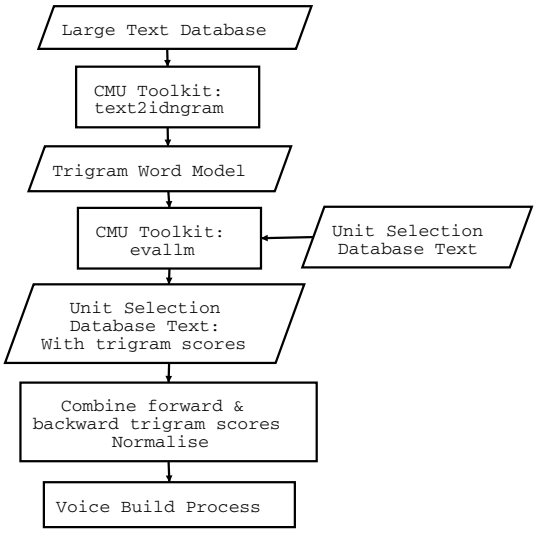

Figure 1: *Process for adding trigram information to a unit selection TTS voice.*

1.1. Feature Extraction

For both redundancy and duration a value is required for each unit in the unit selection database and a reference value for the bias cost function (see below).

The duration feature value was calculated using the phone duration in the database. The reference value was the average duration of this phone type for the speaker.

The redundancy feature value was calculated based on word trigram statistics generated from a large text database (approximately 50 million words) using the CMU language toolkit [9]. This was then applied to the speech database to calculate two probabilities, firstly the probability of guessing a word given the previous two words, and secondly the probability of guessing the word given the two subsequent words. This second, 'backwards' trigram probability was calculated to remove the bias of a left right reading of the text. Both probabilities were logged, added and normalised to a value between 0 and 256 where a high value meant it was highly predictable (see Figure 1). Table 1 shows some example values for a sentence. A value of 127, the mean value after normalisation, was used as a reference point.

Table 1: *Example of redundancy feature values.*

Word	Trigram log prob. left-right	Trigram log prob. right-left	Normalised combined value
they	-5.66200	-0.98100	207
will	-3.30700	-2.75800	211
boycott	-7.36900	-11.06800	121
the	-0.64400	-1.55600	239
vaudeville	-3.23700	-3.25200	208

1.2. Cost Calculation

Given a conventional unit selection system (e.g [8]), it is possible to avoid the selection of high redundancy (i.e predictable) and/or short units by adding appropriate target cost functions. An example of such a function would be to return zero cost if a unit's duration is above an average value, or its redundancy is below a reference value, and a linear weighted cost when duration is below such a mean value, or redundancy is higher than a reference value. When hyper-articulated speech is required (say to increase the focus on a word), the weights would be switched on for these functions (see Figure 2).

However the effect of applying these cost functions on increasing focus is far from transparent for the following reasons:

1. As such selection requires appropriate data to be present in the database we might expect such functions to become saturated when weights go above a certain threshold.
2. It is unclear to what extent either duration or redundancy affect perceived prominence and how they interact with each other.
3. It is unclear what the magnitude of the effect of applying such cost functions are in terms of such perceived prominence.

We address these issues by analysing the effect of altering weights on unit selection, by carrying out a perceptual experiment to ascertain resulting prominence/focus, and by looking at the relationship between the underlying features and the perception of prominence.

2. Synthesis Experiment

Four versions of twenty anomalous haskins sentences were synthesised using the rhetorical system's rVoice speech synthesis system. A word was chosen from each sentences to become focused. Haskins' sentences are of the form:

`The wrong shot led the farm.`

Only the content words could receive focus. Version one was synthesised normally, version two switched on cost function to remove short units, version three applied a cost function to remove high redundancy units, while version four did both.

2.1. Determining the weight of the cost functions

Figure 3 shows how the feature values of the selected units change as the weight of the cost functions are increased (in this case the average duration of phones in the target words). For both cost functions we chose a weight just at the point of saturation. This was to achieve the biggest perceptual change. The weight scale is left unmarked as the value is meaningless except in the terms of altering the viterbi behaviour.

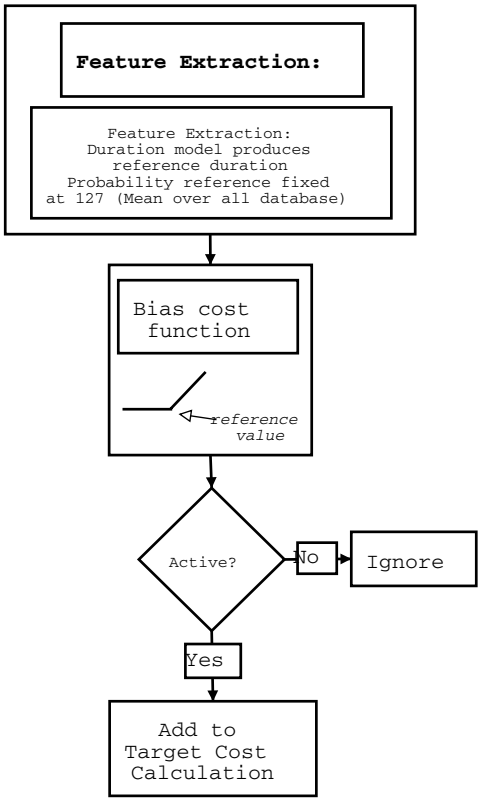

Figure 2: *Modification of target cost during synthesis. The same style of cost function is used both for duration and redundancy but with a different reference points. In addition, for duration a cost is returned as duration falls while for redundancy a cost is returned as redundancy rises.*

2.2. Perceptual Test

- The eighty sentences were divided into 4 sets of stimuli with control, duration, reduction and both evenly distributed across the sets.
- The word chosen to increase focus was also evenly distributed across the four possible content words.
- Each subject listened to only one version of each sentence.
- The voice used was based on Scansoft Realspeak Emily, a British English voice, which had been transferred into the rVoice system.

In English there is a marked default to regard sentencial accent as falling on the last word. In order to cope with this bias, subjects were asked to mark both the primary and the secondary focused words. The subjects could listen to the sentence as many times as they liked and they were shown the transcription.

Figure 4 shows the average prominence assigned to each word. The word position has a significant effect not only on the perceived prominence but also on how effective the cost functions were in raising it.

A paired t-test with bonferroni correction was carried out between the average perceived prominence across subjects for the control sentences and each cost function condition, see Table 2.

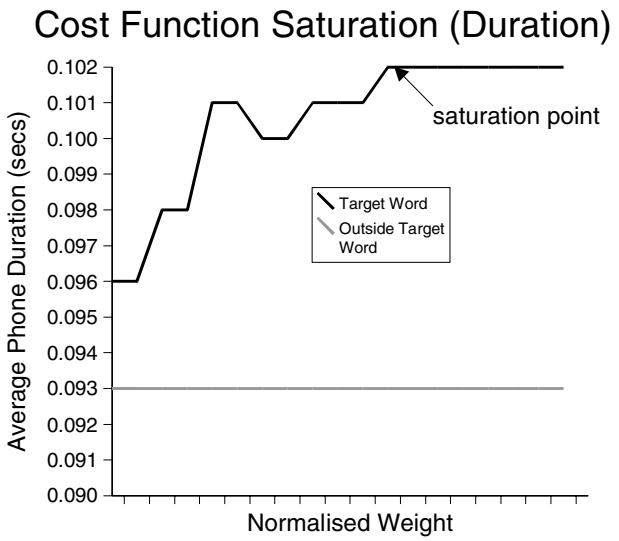

Figure 3: *The feature acted upon by the cost function reaches a saturation point. This was used to set the weight for the cost function in the experiment. The average phone duration of material outside the target effect of the cost function, in the rest of the sentence, is shown as a reference.*

Table 2: *Paired t test results between the control, duration bias, redundancy bias and both cost functions.*

Cost Function	Mean	Control Mean	t	df	sig.
Duration	0.54	0.37	-3.11	19	$p < 0.01$
Redundancy	0.49	0.37	-1.940	19	$p = 0.067$ (NS)
Both	0.51	0.37	-1.83	19	$p = 0.083$ (NS)

The most effective cost function was based on duration alone. However it was only able to raise the chance of a perceived prominence by 17%. This is much less than the basic effect of word position.

3. Relationship between features and perceived prominence

The results from the perception test show that the cost functions were (with limited success), able to increase the perceived prominence of a chosen word. This raises the question of how variation in features affected perceived prominence. A linear regression carried out on the average perceived prominence over the three cost function groups with average trigram redundancy and phone duration factors across target word was non significant. However this regression was heavily confounded by word position. If final word data was excluded the two factors predicted 26% of the average perceived prominence (r 0.512 rsquared 0.26 p$<$ 0.005). We can estimate how independent the factors are by carrying out the regression with each factor removed and noting the value of the drop in predictive power. Then, by subtracting these values from the overall predictive power, we can calculate the non unique contribution from the model. The two factors act quite independently in this regression with a non unique contribution of only 1% (unique contribution from duration - 11%, from redundancy - 14%).

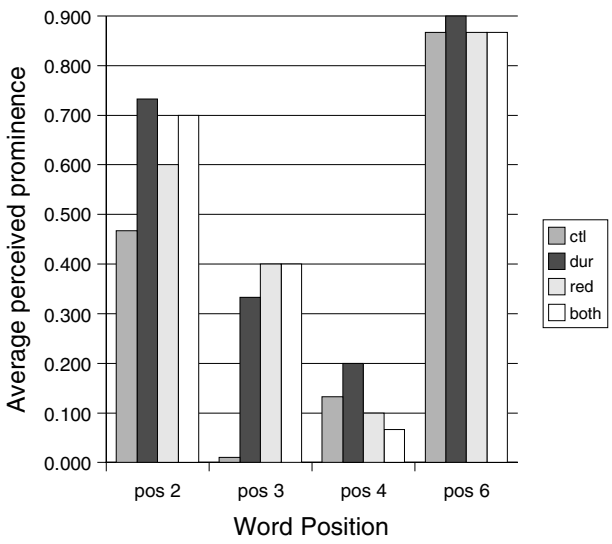

Figure 4: *Average perceived prominence. The words were taken from a haskins six word sentence of the form 'The word2 word3 word4 the word6' (see section 2). Only content words could receive extra focus. The histograms show perceived prominence by word position and by cost function (ctrl - none, dur - duration bias, red - redundancy bias, both - both duration and redundancy bias).*

4. Discussion

The duration cost function shows some success in increasing perceived prominence. However the effect is far from systematic. rVoice implemented these cost functions as xml tags which could be attached to any word in the input. However because of the unreliable nature of the effect of the cost functions they were coupled with other DSP effects (such as a forced increase in duration) in the rVoice specific pros-care tag. In addition these tags could be automatically assigned by the text front end to proper names or words in the lexicon that users wish to be more emphatically stated.

However such an implementation does not offer real control of prominence and focus. There are a number of problems which need to be addressed to give the sort of control some users (or front ends) require for a general unit selection engine.

4.1. Cost function design

The duration cost function has an underlying assumption that increased duration will necessarily increase perceived prominence, and that by selecting long units from the database rather than using post processing to increase the duration the effect will be more natural. In fact unstressed phrase final units are a clear example of units where increased duration does not necessitate an increased perception of prominence. This allows the choice of either preserving phrase position and having less long units to select, or of relaxing phrase position and potentially selecting units which give an unnatural impression of phrasing and may not increase the prominence. Solving this problem requires a more detailed model of duration across phrase and prominence contexts.

The redundancy cost function increased the perceived prominence but not significantly. This could be ascribed to noise in the basic redundancy feature. The combination left-right/right-left word trigram normalised feature may not be op-

timal.

The functions expressly bias only against unwanted units below the reference value. This was to prevent the accidental selection of outliers. However this also contributes to saturation and reduces control.

4.2. Effects on viterbi

By increasing the target cost for such functions you are implicitly reducing the contribution of the transition cost in the viterbi search. The result is a tight rope walk between selecting words with sufficient focus and retaining smooth concatenation. In addition, despite the relative independence of the features to perceived prominence (see section 3), the combination of the cost functions tended to perform worse than either alone. Pre-pruning of tokens before carrying out viterbi could also have a big impact on the saturation of such cost functions.

4.3. Database design

The problem of saturation could be dealt with by expressly designing databases which include hyper-articulated data. However such a brute force process is resource intensive. In addition, it raises the issue of whether preserving word identity should take precedence over these cost functions. For example, should material for the word 'two' come from the word 'tutankamen' when you want to hyper-articulate it? If not then we may not be able to hyper-articulate common words, if yes we might see serious quality degradation caused by unexpected prosody and increased concatenation.

4.4. Perception of Prominence

We have a poor understanding of the perceptual process which underlies the perception of prominence and focus. For example perhaps reducing unfocused words would help shift the perceived focus. We might also use signal processing to increase the perceived effect. Finally it is possible to look at other correlates of focus such as spectral tilt, amplitude or f0 transition.

5. Conclusions

Setting up local cost functions which can be manipulated using xml input tags is a practical solution to modifying the prosodic structure of unit selection synthesisers. However the extent modification occurs is dependent on non trivial interactions.

1. *The interaction between the the contents of the database and the cost function:* If the units with the desired acoustic characteristics are not present they can't be selected.
2. *The interaction between the realised change in the dependent acoustic feature and perception of the effect required:* Many other possible factors can alter the perception of this effect. In this example, word position confounds most attempts to remove focus from the ultimate word in these short sentences.
3. *The interaction between local cost functions:* In this example, although both duration and redundancy cost functions raised average perceived prominence, the combination of the two was less successful than duration alone, despite the apparent orthogonal effect of the dependent acoustic features on perceived prominence. This is likely to relate to the *non* orthogonal relationship between the features values assigned to units in the database.
4. *The interaction between other perceptual effects required and the cost functions:* For example, we may succeed in making the words more focused but at the expense of naturalness.

6. References

[1] M. Anderson, J. Pierrehumbert, and M. Liberman, "Synthesis by rule of english intonation patterns," in *ICASSP*, 1984, pp. 281–284.

[2] J. Fackrell, H. Vereecken, C. Grover, J. Martens, and B. Van Coile, "Corpus-based development of prosodic models across six languages," in *Improvements in Speech Synthesis*, E. Keller, G. Bailey, A. Monaghan, J. Terken, and M. Huckvale, Eds. Wiley, 2002.

[3] K.E. Dusterhoff, A.W. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict f0 contours," in *Eurospeech*, 1999, pp. 1627–1630.

[4] L. Blin and L. Miclet, "Generating synthetic speech prosody with lazy learning in tree structures," in *CoNLL-2000 and LLL-2000*, 2000, pp. 87–90.

[5] J. Meron, "Prosodic unit selection using an imitation speech database," in *4th ISCA Workshop on Speech Synthesis*, 2001, pp. 53–57.

[6] Matthew P. Aylett, *Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech (http://www.cogsci.ed.ac.uk/m̃atthewa/thesis_sum.html)*, Ph.D. thesis, University of Edinburgh, 2000.

[7] Matthew Aylett and Alice Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, pp. 31–56, 2004.

[8] A.J. Hunt and A.W. Black, "Unit selection in concatanative speech synthesis using a large speech database," in *ICASSP*, 1996, vol. 1, pp. 192–252.

[9] Philip Clarkson and Ronald Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit.," in *Proceedings of Eurospeech 97*, 1997, pp. 2707–10.