Prosody and the Selection of Units for Concatenation Synthesis

Nick Campbell

ATR Interpreting Telecommunications Research Laboratories

Hikaridai 2-2, Seika-cho, Soraku-gun, Kyoto 619-02 Japan. e-mail nick@itl.atr.co.jp

Abstract

The ATR ν-talk non-uniform unit system of concatenative synthesis [1] has been shown to produce very high quality synthetic speech, but is slow and expensive in memory. Furthermore, it was designed for Japanese and is not directly applicable to other languages. This paper shows how the ν-talk principle can be generalised for multi-lingual synthesis, and describes methods for database pruning and faster unit selection that overcome the main criticisms levelled against the Japanese version. To reduce selection time, we substitute prosodic selection criteria for the acoustic measures, and show that these result in faster unit selection that minimises post-processing of the speech waveform and thus reduces distortion in the output speech. To reduce database size, we generate a rectangular array of non-uniform segments to a predetermined depth. This preserves sparse units and maximises tokens of the common sounds of the language.

1. Introduction

The relation between the f0 of an utterance and variation in its spectral characteristics has long been known [2], but most concatenative synthesis methods still employ a relatively small fixed number of source units, one token per type, under the assumption that any modification of their inherent pitch and duration can be performed independently at a later stage of processing. The distortion of the resultant speech that is introduced by changing a segment's duration or fundamental frequency has until recently been masked by the generally poor, mechanical quality of the generated speech after it has passed through a coding stage. However, as synthesis quality has improved, and as the memory limitations of earlier systems are eased, it now becomes necessary to reconsider the usefulness of such limited unit sets.

In a typical speech database, the number of types (labels) is small, and the number of tokens varies from very large (for a few vowels) to extremely few (for some rare consonant combinations). Unit inventories extracted from such corpora typically ignore extra tokens of frequently-occurring sounds, but by labelling the prosodic characteristics of the segments in identical contexts, use can be made of these supposedly duplicate units to reduce subsequent distortion.

Non-uniform-unit synthesis uses acoustic measures to search the whole database for concatenation units, rather than pre-select one of each type, but searching for an optimal token is therefore time-consuming. Efficient pruning of the data is required, so to reduce the size of the source database while maximising the variety of units within it, I propose the following steps:

i) Subdivide the frequent units, and cluster them with their most common neighbours to form longer units, As the number of types grows, the number of tokens of each diminishes until the type-token array becomes uniform.

Quartiles of the Euclidean cepstral distance measure

	min	25%	median	75%	max
segmental context alone:	0.0071	0.3965	0.8581	1.7219	8.8546
segmental & prosodic ctxt:	0.0073	0.3167	0.6232	1.4390	10.3748
seg alone after psola:	0.0128	0.2406	0.5838	1.1891	9.1189
seg + pros after psola:	0.0106	0.2218	0.4331	1.0501	7.2118

5. Discussion

At this level of the synthesis process, a database is defined solely by the labels used to transcribe it. The unit-generation algorithm produces a unit set that best models the collocation frequencies of the input data in terms of its own labels. The prosodic information is generated automatically from the speech data, given the label set. The method is thus language independent, and relies only on an adequate size corpus from which to draw the units, and a language interface by which to generate the transcriptions for the utterance to be synthesised. The method is currently being tested with several databases from different speakers of both English and Japanese, under different labelling conventions, and appears immune to language or labelling differences.

The advantage of specifying prosodic variation in terms of variance about a mean, and slope in terms of the differential of the normalised measures, is that regardless of the prediction values, retrieved values are constrained to be within the natural range for the speaker's voice. Describing the pitch of a segment as 'moderately high and rising' ensures that the closest unit in the database will be selected, and in many cases the difference between the target pitch and the unit's original is small enough to be perceptually insignificant. Experience with this system encourages us to believe that in the majority of cases it is better to relax our target goals in the direction of the database events than to impose an unnatural (and possibly distorting) pitch or duration on the waveform.

6. Conclusion

It has been shown that prosodic variation has more than a small effect on the spectral characteristics of speech, and that advantage can be taken of this in the selection of units for concatenative synthesis. It was also shown that a database of non-uniform units can be automatically generated from a labelled corpus and that the prosodic characteristics of contour shape and excursion can be automatically coded. Nothing above will make up for the lack of an appropriate unit in a corpus, and careful design of this resource is essential, but a way of making better use of the supposedly redundant duplicate tokens has been suggested.

References

[1] Tekeda, K., Abe, K., and Sagisaka, Y.: "On the basic scheme and algorithms in non-uniform unit speech synthesis", pp. 93-106 in "Talking Machines,", eds Bailly, G. and Benoit, C., North-Holland, 1992

[2] Traunmüller, H.: "Functions and limits of the F1:F0 covariation in speech", pp. 125-130 in PERILUS XIV, Department of Phonetics, Stockholm University, 1991

[3] ESPS/waves+ , Entropic Research Laboratory, Inc, 600 Pennsylvania Avenue, Washington DC 20003.

[4] Campbell. W. N.: "Syllable-based segment duration", pp. 211-224 in "Talking Machines,", edited by Bailly, G. and Benoit, C., North-Holland, 1992

For each phone sequence in the input string, all tokens of any non-uniform unit that covers that sequence are extracted from the database. These candidate units are first ranked according to length, then scored for segmental context (both exact and broad-class fit), and for prosodic appropriateness. This pre-selection greatly reduces the search space of the original ν-talk system, and replaces the computationally expensive acoustic measures of goodness-of-fit.

In the selection, equal weighting is given to the three criteria (segmental and prosodic context, and unit length) by ranking the candidate tokens according to the scores in each, and selecting the top n candidates from the sum of the ranks. A final stage evaluates the optimal path through the candidate tokens by tracking coverage of each unit forward through the target utterance. If a later candidate has a better prosodic score at any point, it overrides the current unit, replacing it to guarantee continuation under the best prosodic conditions available from the database. In this way, the length of a non-uniform unit becomes a weaker criterion than its prosodic suitability.

4. Quantitative evaluation

To determine whether the inclusion of prosody is justified as a selection criterion, the following test was performed. The units for our synthesiser are selected from a database of 503 magazine and newspaper sentence readings. Each in turn was excluded from the database, and an equivalent utterance generated from the remaining tokens. The resynthesised version was then compared with the original, using measures of cepstral similarity. Since the cepstral measure, widely used in speech recognition, is sensitive primarily to spectral features of the speech, it serves well to confirm any effect of prosodic variation on spectral colouring. Comparisons were made between the original recording of each sentence, and resynthesised versions with and without prosodic selection. They were performed first on raw output and then on PSOLA modified output warped to match the prosody of the target utterance.

The source database (26,264 phone segments) yielded 70 original labels (including segment clusters that could not be reliably separated), which after processing formed 635 non-uniform units ranging in length from 1 to 7 original labels. It was pruned to a maximum depth of 35 tokens each. Labels and raw prosodic values (duration and mean pitch and energy for each phone) for a test set of 100 randomly selected sentences were extracted, then the original sentence data was removed before resynthesis. Non-weighted Euclidean measures of the cepstral distance between the original utterance and each resynthesised version were calculated on a phone-by-phone basis after LPC cepstral coding of the waveforms using the default settings of HCode (HTK [3]) to produce 12 coefficients per 10 msec frame.

Results showed that an improvement was gained by inclusion of prosodic information in the selection (*seg only vs. seg+pros:* $t = 4.484, df = 6474$), and a further improvement gained after PSOLA modification of pitch and duration to match the target (*before vs after modification:* $t = 8.312, df = 6474$). Both results are significant at $p < 0.001$. The inclusion of prosodic selection resulted in a decrease in distance from 0.86 to 0.43 (median values). This can be compared with a median distance of 0.58 for identical PSOLA treatment of the concatenated segments selected without consideration of their prosodic environment.

ii) Label each token according to its prosodic characteristics and quantise them to maximise the difference along the prosodic dimension for each type.
iii) Specify a depth n to which to limit the search, and select the n most prosodically diverse tokens for each type, then remove any remaining tokens from consideration.

The clustering grows units that tend to approximate the most common collocations in the source data, and thus models e.g., the frequently-occurring and frequently-reduced function-words automatically. The initial pruning according to prosodic criteria ensures that the tokens remaining in the database are maximally representative of the variation in the speech. Rare tokens are automatically preserved. The choice of an appropriate value for n is then a trade-off between compact size and output voice quality; a larger database is more likely to contain a prosodically appropriate segment that will need less modification to reach a target setting in the concatenated utterance.

2. Prosodic labelling

This section describes prosodic labelling for both database construction and unit selection. It argues that prosodic events are better characteristised with relative values, as these normalise for both speaker individuality and utterance conditions.

The functions of prosody are diverse, but at the simplest level we are concerned with marking prominence and delimiting phrasing in the speech. An experienced transcriber can recognise these features in an utterance from the shape of the fundamental frequency contour and from differences in the lengthening and amplitude of the waveform segments. When describing a contour for generation, it may be more appropriate to specify simply that a unit should be 'long', or 'from a rising part of the contour'.

In order to model prosodic events independently of absolute values, each parameter is normalised twice; first to express its value as a position within a range, and then as a part of a contour. For this, two functions from the ESPS waves+ program can be used (ver 5.0 [3]). The first *'get_f0'* gives an estimate of the fundamental frequency, probability of voicing, zero-crossing rate, and rms energy for every 10 msec of speech in the database; the second *'stats'* gives summary statistics of these for each segment described by a (phone) label in the database. These values are then z-score normalised for each phone class [4] to express the difference of each segment from the mean for its type (in terms of the observed variance for other phones of that type) for the three dimensions of pitch, energy, and duration. To model the position of each token in relation to the respective prosodic contours, first differences of the normalised values are then taken over a window of three phones to the left and right of each segment. The sign of the result indicates whether a segment is part of a rising contour (increasing pitch, loudness or length) or falling. The magnitude indicates the rapidity of the change.

3. Prosody-based unit selection

Each segment is thus assigned six values to describe its position with respect to the prosodic contours of the utterance; in terms of excursion from the mean and position in the contour, for the three prosodic dimensions. When selecting units thus encoded, specification of target prosody in terms of z-scores ensures that the output speech is naturally constrained to be within the range of the source speaker.