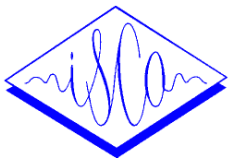

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Duration, pitch and diphones in the CSTR TTS system.

*W. N. Campbell, S. D. Isard, A. I. C. Monaghan & J. Verhoeven

Edinburgh University, Centre for Speech Technology Research.
80 South Bridge, Edinburgh, EH1 1HN UK.

*Now at ATR Interpreting Telephony Research Laboratories
Sanpeidani, Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

This paper describes the prosodic processing and waveform generation components of the text-to-speech system being developed at Edinburgh University's Centre for Speech Technology Research.

Intonation is specified as a sequence of minimal descriptors whose locations are given in terms of syntactically-determined prosodic domains. A pitch contour is computed by converting the descriptors into a sequence of abstract targets whose absolute values depend on a specific speaker model. Duration is determined first at the level of the syllable by a neural network, then accommodated at the segment level according to the distributions observed in a phonetically balanced database. The output waveform is generated by LPC resynthesis of diphone units. Three methods of diphone segmentation are discussed.

1 Introduction

Phonological theories of intonation and duration are being tested in the construction of a text-to-speech system.

A minimal input specification language describes the key prosodic events in abstract terms and broadly describes the phonemic sequence of the text to be spoken. From this, pitch and duration parameters are calculated and the output waveform generated.

In order to minimise the distortion inherent in the waveform-generation component of the system, diphone concatenation has been adopted for the modelling of the voice. Three methods of diphone segmentation are currently being tested:

1. each diphone being assigned fixed endpoints in acoustically stable regions.
2. each diphone being assigned fixed endpoints in such a way as to minimize cepstral mismatch when diphones are joined together.
3. each diphone being assigned a full set of possible endpoints, corresponding to all of the other diphones that it might join to, in such a way that the cepstral mismatch at each individual diphone join is minimized.

Once pitch ranges and targets have been determined for each intonation group, duration is calculated for each segment in the utterance, and diphones are resynthesized with the specified intonation and duration. They are selected from a library of approximately two thousand units, taken from recordings of phonetically balanced nonsense words such that every valid combination of phones in the English language is provided for.

2 Input to the system

Plain text input is preprocessed by the text-analysis component and arrives in the form of machine-readable phonemic characters interspersed with diacritics specifying word stress, pitch accent location and boundary information.

Phonetic diacritics attached to the machine-readable segment labels include [$] for word-edge allophonic variants, and [:] for clustered consonants. The remaining diacritics shown here represent boundaries ([.], [—], [+] and [$] for syllable, word, phonological and intonational phrase respectively), stress ([-], [ ] and [*] for unstressed, reduced, and stressed syllables respectively), and accent ([1], [2] and [3] for primary, secondary and tertiary accents respectively).

By way of an example, sentence No 28 in the SCRIBE phonetically balanced database reads: *The mud squelched loudly and he realised that his suede boots were doomed.* After analysis and transription this is received in the following form:

```
- dh @ | * 2 m uh d | * s: :k w e ll ch t | *
1 l au d . l ii  + | - @ n d | - h ii | * 3 r
i@ . l ai z d + | - dh @ t | - h i z | * 2 s:
:w ei d | * b uu t s | w @ | * 1 $d uu m d $|
```

Figure 1: Input to the components

The first word, *The*, being a function word and therefore presumably reduced, is indicated as such by the [-] preceeding it. The second, a noun, is both stressed (*) and carries a secondary accent (2: non-nuclear, non-continuation). *Squelched* here is marked as stressed but with no pitch prominence; *loudly* carries the nuclear stress and is followed by a phonological phrase boundary marker (+). *doomed* by an intonational phrase boundray marker, and so on. All such prosodic tags are interpreted for both pitch and the durational consequences. Accommodation of phonetic effects such as assimilation and elision is inherent in the diphone principle and does not need explicit specification in the input string.

3 Duration determination

The two main theoretical assumptions of the duration model are

1. That speech timing is a function of higher-level processes operating at the level of the syllable and above.
2. That segment duration can be determined by accommodation into a syllable frame according to known probability distribution functions for each phoneme.

The durational characteristics of each syllable are first determined by consideration of its status in terms of stress, position in the phrase, phonetic complexity, and position in the rhythmic sequence. A neural net trained on a corpus of such syllable-feature – duration pairs, from a twenty-minute passage of broadcast speech, predicts the overall syllable durations. When tested with three different thousand-syllable passages from the same corpus as the training set, this method accounted for approximately 70% of the variance observed in each of the test cases [1] without taking variation of global speech rate into account.

Segment duration is calculated as a secondary process, distributing the predicted syllable duration among the component segments according to their observed distribution densities. Means (μ_i) and standard deviations (σ_i) for each phone$_i$ are computed in the log domain from segment measurements of a phonetically-balanced 200-sentence corpus of read speech. Given an n segment syllable which has been assigned a duration of Δ by the neural net, we solve following formula for k:

$$\Delta = \sum_{i=1}^{n} exp(\mu_i + k\sigma_i) \qquad (1)$$

Segment$_i$ is assigned the duration $exp(\mu_i + k\sigma_i)$. For further details see Campbell 1990 [2] (this volume), and Campbell and Isard [3] (in press).

Further to the example above, the durations produced by this component are detailed in Fig 2 below for comparison. It will be noticed that some editing has been required to match the segment string to the spoken form as transcribed; the present form of the text analysis component of the system does not perform the /h/-dropping that is common in fluent natural speech, much of which is recorded in the diphones by default. TTS output is shown above, natural durations, taken from the original reading in the corpus, below (values in milliseconds):

```
-------------------------------------------------
dh 32 @ 48 m 75 uh 104 d 61 s_  77 _k 65 w 38 e  69
dh 33 @ 37 m 81 uh 110 d 74 s_ 116 _k 56 w 35 e 109

ll 43 ch 106 t 44 l 49 au 163 d 45 l 64 ii 119 @ 31
ll 44 ch  94 t 41 l 59 au 205 d 40 l 65 ii 158 @ 52

n 56 ii 78 r 60 i@ 156 l 46 ai 122 z 68 d 43 dh 24
n 41 ii 89 r 84 i@ 159 l 53 ai 179 z 86 d 13 dh 13

@ 38 t 55 i 42 z 56 s_ 101 _w 52 ei 128 d 45 b 66
@ 47 t 39 i 74 z 68 s_ 109 _w 39 ei 144 d 69 b 75

uu  90 t 59 s  93 w 59 @ 50 $d 65 uu 154 m 97 d 119
uu 140 t 60 s 145 w 33 @ 46 $d 88 uu 173 m 137 d 75
-------------------------------------------------
```

Figure 2: Duration output

It can be seen that the model produces less extreme values than the human speaker (GSW), but it has access to far less information regarding the interpretation of the text. However, it does appear from these results that the model is sufficient for our present purposes. Further research is needed to determine the type of variation required for changes in speaking rate, which is currently implemented by modification of the parameter k in the above formula.

4 Pitch processing

The intonation model implemented in the CSTR system is similar to Pierrehumbert's target-and-transition approach [6]. To generate a contour from a set of specified pitch accents requires two stages of processing: interpreting accents as phonological targets, and calculating absolute phonetic values for those targets. Targets are currently linked by straight-line interpolation. The adequacy of the target-and transition approach and the accuracy of our phonetic model are therefore crucial for the system's performance.

4.1 Target assignment

The interpretation of accents and boundaries as targets involves reference to a 'tune', which defines the type of utterance (and thus the type of contour) to be produced. Tune choice depends on the speech act conveyed by the text, and is not currently implemented: instead, there is a default tune which

produces satisfactory intonation contours for most declarative and WH-question utterances in English.

The tune specifies the type of primary accent, the type of secondary accent, and the initial and final boundaries to be assigned: it is assumed that tertiary accents have an invariant interpretation. The default tune specifies a H*L primary, a H* secondary, a mid initial boundary and a low final boundary: tertiaries are seen as H*H.

> H* is interpreted as a mid target followed by a high target, i.e. a rise.
>
> H*L is interpreted as a mid target followed by a high target followed by a low target, i.e. a rise-fall.
>
> H*H is interpreted as a mid target followed by a high target followed by another high target, i.e. a sustained rise.

The first two targets in an accent are placed one segment before and 60% through the accented vowel respectively: the third (or trailing) target of a primary accent is associated with the following vowel. The trailing high target of a tertiary accent is postponed until the next accent, thus linking the two accented items: tertiary accents encode close dependency on the following accented item. Elements which have not been accented at any stage are marked with a hyphen [-] and currently have their lexical stress marks removed as they do not constitute prominent elements in any sense.

4.2 The phonetic model

The phonetic model is much as described in [4] and [5]. The interpretation of targets is based on a phonological REGISTER which defines high, mid and low lines. High and low are similar in some ways to top and bottom declination lines in other models: mid is a neutral line to which the contour tends to return after pitch excursions. This register can approach the speaker baseline (downstep) or move away from it (upstep). In our model there are speaker-specific register parameters (minimum Fo, register width, initial height of register relative to minimum Fo) and speaker-independent parameters (excursion size relative to register width, step height relative to register height, current register height). These parameters are all inherently relative, with the possible exception of the speaker-specific parameters.

The interpretation of boundaries by the phonetic model involves three parameters which constitute the phrasal component in the Fo equations. They are a tg(0) or phonological phrase parameter, a tg(1) (intonational phrase) parameter, and a tg(2) (sentence/utterance) parameter respectively, all of which are equal to 1 by default. At any tg(X) boundary, the default treatment is to downstep (reduce, or multiply by 0.8 in the current system) the tg(X) parameter and reset any tg(Y) parameters (where Y < X) to 1. Thus, successive tg(0) boundaries will create a decline in Fo which is arrested when the tg(0) phrasal parameter is restored to its initial value at tg(1) boundaries: similarly, successive tg(1) boundaries will progressively lower Fo until a tg(2) boundary resets the tg(0) and tg(1) phrasal parameters to their default values. This produces a natural-sounding contour in which both hierarchical and local relations can be expressed in a regular manner. Since the default is to downstep at every boundary, an approximation to declination models [8] [9] will generally be produced - however, there is nothing to prevent our model upstepping at certain boundaries and this allows us to produce more natural-sounding output than a standard declination model.

No manipulation of other linguistic variables (such as excursion size) is currently undertaken, as the information which controls such variation is not yet available from text. It is expected that this will change in the near future.

5 Diphones

The basic idea behind diphone synthesis is that although segment boundaries, insofar as they can be said to exist at all, are extremely variable and context dependent, the middles of segments often contain acoustically stable regions which do not depend as much on the identities of neighbouring segments. Diphone synthesis joins together bits of natural speech stretching from the middle of one segment to the middle of the next and the result can be highly natural and intelligible, with the joins between diphones not too intrusive.

The basic expectation underlying diphone synthesis, that diphones will join together smoothly, has to be regarded as an illusion though, in that intrinsic phonetic variability and slight variations in the recording conditions can cause differences in the realization of the articulatory targets for each phoneme, which is reflected in the spectral characteristics of the diphones. As a result, discontinuities in amplitude and spectrum between abutting diphones in speech synthesis are unavoidable and can potentially degrade the system's performance by introducing audible clicks and a general impression of roughness.

As an alternative to the traditional method of hand-segmentati two further diphone segmentation methods have been developed which aim to obtain optimally segmented diphones, i.e. diphones which have been extracted in such a way that minimal spectral discontinuity between all possible pairs of diphones is guaranteed a priori. The performance of these methods has been assessed with respect to the hand-segmentation method.

In the first method, an 'average' ceptrum for all tokens of, say /e/, is calculated and each individual /-e/ or /e-/ diphone is segmented at a point where its cepstrum is as close

as possible to the overall /e/ average. This technique was used to segment vocalic regions only; the cepstra of fricative regions are too variable and such sophistication is wasted on silent regions.

With this method, an overall reduction of spectral discontinuity between all hypothetically possible pairs of diphones of 11% was achieved as compared to the same diphone combinations from the hand-segmented inventory, which thus constitutes a slight improvement. It was however found that this gain is more apparent than real in that this method actually increases spectral discontinuity in about half the cases. In the other half, discontinuity improves in such a way that the average gains slightly outperform the losses. As a result, a diphone inventory extracted by this method cannot be considered as an improvement in real terms.

In the second method, the idea of using fixed boundaries for each diphone was abandoned and the concept of diphone context-sensitivity was explored. Context-sensitive diphones are essentially flexible units in that they have variable boundaries which are made dependent on the diphones they combine with in the synthesis process. At the point in the synthesis stage when it is known which sounds are required to produce an utterance, the optimal joins of the diphones in the utterance are determined by an algorithm that aims to reduce spectral discontinuity between each diphone and its adjoining diphones. This method involves storing much more information than either of the other two schemes but it produces a decrease in cepstral mismatch of between 55% and 81%, and workers accustomed to listening to the system can hear the improvement.

The method of context-sensitive diphones was used to establish a complete diphone inventory for two male speakers of British English and formal tests with naive listeners are presently being carried out.

6 Conclusion

Three components of the output stage of a text-to-speech system currently under development at Edinburgh have been described. After text pre-processing, duration and pitch values are determined and diphones concatenated to produce the output waveform.

As a research tool, this text-to-speech system serves to test phonological theories related to both pitch and duration by allowing fast generation of speech waveforms from a minimal specification of the linguistically relevant parameters. By use of abstract representations in the input text, we are able to focus on the linguistic features that correlate most strongly with prosodic variations in the waveform, and with the use of diphone synthesis, we are able to listen to and perceptually evaluate the effects of changes in the techniques of realisation of these variations with minimum interference from the waveform generation component.

Formal tests to evaluate the performance of each of the three components, both separately and in conjunction with each other, are now under way.

Acknowledgements

The authors are particularly grateful to Richard Caley for his work in unifying the system and acknowledge that without his efforts we would be unable to hear the results of ours.

This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech Technology Demonstrator Project (SERC grants D/29604, D/29628, F/10309, F/10316, F/70471) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.

References

[1] W. N. Campbell (1990) *Analog I/O nets for syllable timing* in Speech Communication: Special Issue on Neural Nets and Speech, vol 9 no 1, Elsevier Science Publishers B. V. (North Holland).

[2] W. N. Campbell (1990) *Evidence for a syllable-based model of speech timing* this volume.

[3] W. N. Campbell and S. D. Isard (1990) *Segment Durations in a Syllable Frame* Journal of Phonetics: Special Issue on Speech Synthesis (forthcoming).

[4] D. R. Ladd (1987) *A Model of Intonational Phonology for Use in Speech Synthesis by Rule* in Proceedings of Eurospeech 1987, vol. 2 pp. 21-24. Edinburgh: CEP.

[5] A. I. C. Monaghan & D. R. Ladd (1988) *Speaker-Dependent and Speaker-Independent Parameters in Intonation* in Proceedings of ESCA Workshop on Speaker Characterisation, Edinburgh, June 26-28 1990, pp. 167-174.

[6] J. B. Pierrehumbert (1981) *Synthesising Intonation* JASA 70, pp. 985-995.

[7] M. Stella *Speech Synthesis* in F Fallside & W Woods (eds.) Computer Speech Processing, London:Prentice Hall, 1985, pp. 421-460.

[8] N. Thorsen (1985) *Intonation and Text in Standard Danish* JASA 77, pp. 1014-1030.

[9] N. Willems, R. Collier & J. 't Hart (1980) *A Synthesis Scheme for British English Intonation* JASA 84, pp. 1250-1261.