# Processing a Speech Corpus for CHATR Synthesis

Nick Campbell
ATR Interpreting Telecommunications Research Laboratories
Hikari-dai 2-2, Kyoto 619-02, Japan.
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

## Abstract

This paper discusses multi-level labelling of sounds in speech and reviews the segment selection process in the CHATR speech synthesis system [2]. It shows how the system can be refined so that all the information required for the generation of a meaningful utterance can be marked on the source data alone. By this concept of 'intelligent data' we leave only the tasks of index-based retrieval and simple waveform concatenation for the synthesiser, which thus becomes a fast and flexible speaker-independent and language-independent utterance generator.

## 1 Introduction

In a conventional concatenative synthesis system, we find modules for text processing, for prosody prediction, and for signal processing. Text taken as input is converted into a sequence of phonemes for which an 'appropriate' set of prosodic contours (fundamental frequency, duration, and amplitude) is calculated, and then short speech waveform segments, typically stored as diphones or demi-syllable units, are modified to suit the prosodic requirements before concatenation for output as a speech waveform.

CHATR[3] differs from the above in two significant ways: it relegates the waveform store to an external corpus, and it uses the natural variation of the source units to reproduce the desired prosodic characteristics in the synthesised speech. In so doing, it removes the necessity for signal processing (and thereby minimises the distortion in the resulting speech) but instead requires a larger library of source units, and the creation of an index into 30,000 to 50,000 phoneme-sized segments, rather than to 2,000 diphones.

### 1.1 Selection of segments

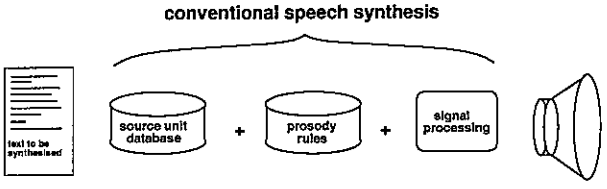



Figure 1: Many existing speech synthesisers use a diphone database and require signal processing for subsequent modification of prosody

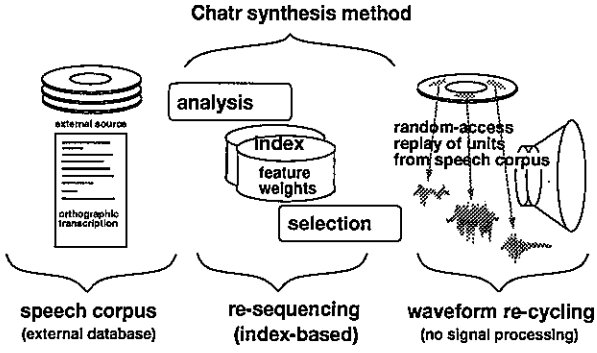
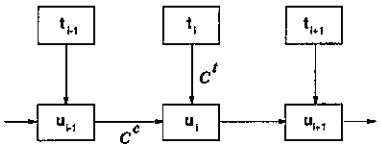



Figure 2: CHATR uses an external speech database and selects segments by prosodic as well as phonemic context for simple re-sequencing.

Figure 1 shows the sequence of processing in a traditional system. Figure 2 shows how CHATR eliminates much of this work, at the expense of the pre-processing required to provide an index into the external speech corpus. The benefit of the large-corpus approach to speech synthesis is that it captures the voice quality and speaking-style characteristics of the original speaker and, given that memory and storage are becoming cheaper and faster every year, offers a simple way to customise and change voices.

In this paper we describe a further simplification to CHATR that removes the intermediate processing and reduces the synthesiser to a simple waveform retrieval device. Because we cannot predict waveform characteristics directly, we access them instead by a feature matrix of their segmental and prosodic characteristics using an index. Previous versions used phone-sized waveform segments, selected to maximise their prosodic and phonemic fit to a desired target utterance by minimising two costs:




*Minimising two distance measures*

The target cost, $C^t(u_i, t_i)$, represents the difference between a database unit $u_i$ and a target $t_i$, calculated as the weighted sum of the differences between the elements of the target and candidate feature vectors. In the current system we use about $p = 30$ target sub-costs. The target cost, given weights $w_j^t$ for the sub-costs, is expressed as

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i)$$

The concatenation cost, $C^c(u_{i-1}, u_i)$, represents the smoothness between consecutive units ($u_{i-1}$ and $u_i$), determined by the weighted sum of $q$ concatenation sub-costs, $C_j^c(u_{i-1}, u_i)$ $(j = 1, ..., q)$. If $u_{i-1}$ and $u_i$ are consecutive units in the synthesis database, then their concatenation cost is zero.

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i)$$

The best combination of units $\overline{u}_1^n$ from those available in the database is found by minimising the total cost $\overline{u}_1^n = \min_{u_1,...,u_n} C(t_1^n, u_1^n)$ for a sequence of $n$ units:

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) + \sum_{i=2}^{n} \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i)$$

The paradox here is that a 'suitable' prosodic contour has to be predicted first in order to pre-select the candidate units that are closest and that will concatenate to form an utterance with the desired prosodic 'meaning'. However, in order to predict such a contour we use statistical models trained using the features of the same source database in a pre-processing stage. We thereby introduce two sources of error: prediction error and selection error.

### 1.2 Compounding of errors

Taking the case of segmental durations, for example, the significant factors used to predict a duration for synthesis are the nature of the phone, influence of neighboring phones, position with respect to prosodic boundaries, and part-of-speech of the parent word (about 6 factors in all) [7]. Similar factors are used for the prediction of segmental amplitude. However, an error in prediction of 20ms or 2 dB is not uncommon. When such a duration estimate is used as a prosodic target in the selection of units by CHATR, a similar variance of 20 ms between desired and actual durations for the selected units is not uncommon. If these two sources of error were to compound, then the overall difference could be more than the duration of the segment itself.

Similarly, average errors of about 3 semitones [6] have been reported for the prediction of a fundamental frequency contour ($F_0$) for synthesis. Figure 3 illustrates the dangers of this type of compound error. The predicted contour is close to the ideal, and the selected segments are close to the predicted, but in some places the error is additive, with the result that in this example the sentence focus might be perceived as being on *words* rather than on *group*.

### 1.3 Features on the data

A solution to this problem could be engineered by increasing the weights $w_j^t$ for the target sub-costs, but this would only be at the expense of smoothness in the resultant waveform. A better solution can be found when we realise that the factors used for training such prosodic prediction models are typically as simple as the ToBI parameters [1] that are already being used to label the prosodics of the source corpora.

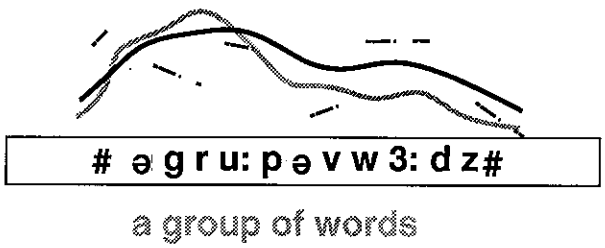



Figure 3: Three contours for a given utterance: grey-line=ideal; solid=predicted; dashed=selected.

Since with CHATR we are selecting segments for synthesis from a natural corpus using segmental context as one of the criteria, then the conditions having the biggest effect on duration and amplitude are automatically satisfied. If we extend the selection to be sensitive to ToBI-type parameters as a specification of the prosodic context, then we reduce the probability of difference between targets and selected candidates by eliminating the prediction stage with its consequent errors.

This elimination of the target prosody contour prediction stage is a logical step in view of the circularity of the train-predict-select sequence. It made sense to predict durations (and $F_0$ and power) when these parameters were used for the subsequent modification of diphones (or non-uniform segments) by signal processing, but now that we have a bigger and more representative selection of prosodic variety in the source units, it makes more sense to select units according to their native prosodic context rather than by proximity to (probably inaccurate) predicted contours.

Instead of using the higher-level features of a target utterance to predict a set of target prosodic contours to restrain the selection of suitable synthesis units, we should use those features directly as unit selection criteria from the database. However, if we are to select units *directly* according to prosodic as well as segmental features, then we must constrain the number of types, and create an index in such a way that these features are as easily accessible as the raw values for duration $F_0$ and power were for the individual phones.

## 2 Multi-level unit selection

Phone-based unit selection offered the advantage of uniform-sized database segments that could be predicted from a dictionary or by letter-to-sound rules and that could easily be aligned to the speech waveform with only minimal manual intervention and checking. It was reasonable to assume that an hour of sampled speech would contain at least one example of each phoneme of the language, and a large number of the typical frequent collocations. Using phone-sized waveform segments we had the advantage of being able to synthesize very natural-sounding sequences of the more common collocations such as verb endings and phrase-particles, and were

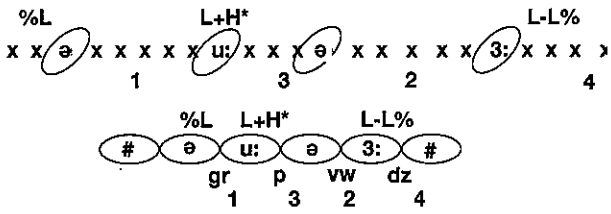



Figure 4: ToBI Prosodic parameters, seen in relation to aligned waveform segments and with syllables for comparison

able to re-construct missing phone sequences by using sub-optimal but close equivalents.

However, as figure 4 shows, the disadvantage of the phone-based approach lies in its irregularity of context with respect to significant prosodic events, which are primarily at the level of the syllable and above. In English, there can be up to seven intervening consonants between a pair of vowels, and so a variable-length window is needed to check intervocalic dependencies. For each individual index entry, we should take into account a previous/following context of at least five phones in order to detect prosodic context. For example, with $F_0$, if the consonants next to a vowel are voiceless, then it makes no sense to measure continuity of fundamental frequency. Selection weights must be conditional and higher-level generalities are not easily trained.

## 2.1 Syllables and sub-phones

Figure 5 shows a solution to this difficulty; linking sub-phonemic waveform labelling (which offers the benefit of finer units for smoother concatenation) with syllabic indexing (which reduces the size of the waveform inventory). Phoneme keys are still predicted using a dictionary but they are mapped down onto sub-phonemic (edge and centre) units for waveform labelling, and up to syllabic (vocoid and contoid pair) labels for the main index.

By labelling speech at the level of the syllable we achieve a uniform unit that is both prosodically and segmentally relevant, enabling a smaller previous/following window requiring only two units of context either side. Every unit now has a measurable fundamental frequency and it becomes relevant to ask whether each unit or its neighbour is accented, whether it has a high or a low tone, and if adjacent to a phrasal boundary, whether it is devoiced or exhibits creak. By asking such questions about only one or two previous and following units we can gain enough knowledge about the context of articulation to be able to 'predict' with some accuracy the manner of production, degree of lengthening, loudness, tonality, and voice-source characteristics, etc.

By encoding the speech database as a sequence of syllables, its distinguishing features can be easily and quickly accessed for the selection of synthesis segments. However, whereas the number of phones in a language may be about 40, the number of possible syllables can be several orders higher. In order to ensure maximal coverage of selection units in an external speech corpus (one not specifically designed for

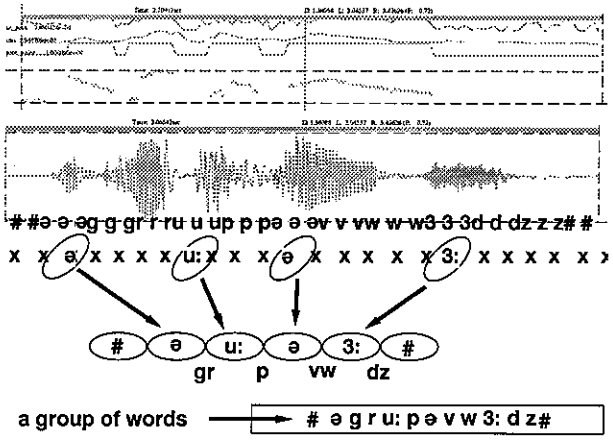



Figure 5: 'A group of words', showing the speech waveform, prosodic vectors, fine segmentation, syllabic segmentation, lexical units and phonemic segmentation.

synthesis purposes) we either need to collect a much larger amount of speech, or to use smaller units.

## 2.2 Features and syllables

Following Öhman [8], we view the main 'carrier' of speech as the vocalic stream ($v\prime$), interspersed with distinguishing consonantal information ($c\prime$). We can thus parametrically encode (or index) waveform characteristics by using a binary bit-vector having the granularity of the syllable. We encode two tiers of phonation, each having effects on the other, but since the prosodic environment has stronger relations with the vocalic tier, it is the syllabic peaks ($v\prime$) that form the core of our index and carry information relating to the syllable as a whole.

For waveform concatenation we only need to locate appropriate 'centres' in each tier, rather than positing absolute segment boundaries. We join between centres at the point of minimal discontinuity in the overlap of their transitions. Since in the best case the transitions out of the vocoid centre will be identical to the transitions into the contoid centre (and vice versa) these joins should be imperceptible.

To reduce the number of unique types, for feature sharing, we maximise similarities in articulatory feature-space and emphasise differences due to prosodics. For example, acoustically similar doubled vowels and geminate consonants can be clustered and distinguished by their durations. The nasal vowel and the semi-vowels, like palatisation or lateralisation in English, or lip-rounding, can be better treated as articulatory features on the pure vowels (again, distinguished also by lengthening). Similarly, devoicing of both vowels and consonants is better treated as a feature of articulation, preserving their place and manner similarities, rather than by labelling the devoiced variants as separate phonemic types.

The syllable peak ($v\prime$) is well described in low-dimensional space (see the IPA vowel triangle, or F1/F2 formant plots for example) but requires annotation as described above for a full specification of vowel quality. Loudness, duration, $F_0$, and spectral-

tilt are not features to be labelled on the syllable per se, but can be predicted from the prosodic environment, which in turn can be largely determined from another bi-level system of peaks and troughs: the prominences and accents marking the focal structure of an utterance, and the phrase and clause boundaries delimiting its chunks.

The contoid tier (*ct*) is also well described by a small number of features like 'strength of intrusion' (weak: approximants, medium: fricatives, strong: plosives), and 'place of articulation' (front:labial, mid:palatal, back:velar), but subject also to influence from the vocoid tier and its prosodic modulations.

Since we encode the two interacting tiers as a sequence of syllable entries in the main index, it is only necessary to characterise each 'syllable' by a *vt* – *ct* pair. Speech is represented in the index as a sequence of two-part bit-vectors describing syllables (starting with a silence syllable) such that the onset characteristics of each subsequent syllable can be derived from the *ct* of the previous.

Unit selection takes place as above, but now without having to force a match between discrete phonemic labels in the initial pre-selection of candidates, and with the advantage of generalisation across features so that speech segments that differ in only insignificant features can be selected when an ideal token is missing from the source corpus.

# 3 Conclusion

This paper has described a new form of multi-level labelling suited to preparing a new voice in the ATR CHATR speech synthesis system. It has the benefit of allowing direct selection of units without requiring an intermediate prediction of numerical targets for prosodic characteristics in the synthesised speech. Rather than predict a numerical target, using models trained on the characteristics of the language as represented by a large corpus of natural speech, we prefer to use the whole corpus for synthesis, after labelling the speech at the level of the syllable. Since certain features, such as boundaries, syntactic class, and prosodic salience are known to be significant predictors of prosodic characteristics, we prefer to use them directly as selection criteria on the units themselves.

Two of the major uses of prosodic information in situations of communication are to encode salience and segmentation [5]. Prominence is a characteristic of the syllable peak, boundary characteristics are a feature of its edges. These correspond to the contexts that are marked by the tone and break-index tiers of a ToBI prosodic annotation [1], which (at least for Japanese) can be performed to a large extent automatically [4].

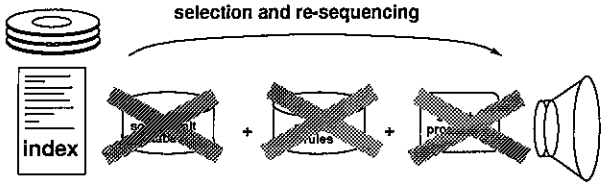



Figure 6: Removing the last stage of processing from the synthesis system. With intelligence in the data, all we need for synthesis is fast selection and re-sequencing.

We have found that in order to predict and replicate the fine details of phonation that characterise situated human speech, it is not necessary to attempt a narrow allophonic specification, but rather that it is better to derive such lower-level details implicitly, based on a higher-level specification of the contexts of the speech situation. Thus, rather than have to predict the micro-melodies and allophonic details of the speech waveform for synthesis, it is sufficient instead to select a sequence of sounds with respect to the two orthogonal spaces of segmentation and salience.

Finally, because the method is now completely language-independent, we are able to model speaker-individuality, dialect, and language differences just by changing databases. We have now created voices for synthesis in five languages, including Kansai Japanese, Tokyo Japanese, Chinese, British and American English, Korean, and German. The syllable-level approach appears to be better suited to all the languages we have studied so far and offers a reduction in the size of the index at the same time as giving easier access to the formative prosodic information.

# References


[1] M. E. Beckman and G. M. Ayers, "The ToBI Handbook", Tech Rept, Ohio-State University, U.S.A. 1993.

[2] W. N. Campbell and A. W. Black, "CHATR: a multi-lingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, (Japanese) 1996(5).

[3] W. N. Campbell, "CHATR: A High-Definition Speech Re-Sequencing System", Proc 3rd ASA/ASJ Joint Meeting, 1223-1228, Hawaii, 1996(12).

[4] W. N. Campbell, "The ToBI system and its application to Japanese" pp223-229, Journal of the Acoustical Society of Japan 53, 3, (in Japanese) 1997.

[5] A. Cutler & D. Norris, "Prosody in situations of communication: salience and segmentation", pp 264-270, Proc ICPhS, Aix-en-Provence, 1995.

[6] T. Hirai, N. Iwahashi, N. Higuchi & Y. Sagisaka, "Automatic extraction of fundamental frequency control rules using statistical analysis", IEICE D-II Vol J78, pp 1572-1580, 1995.

[7] N. Kaiki, K. Takeda, & Y. Sagisaka, "Control of segmental duration for speech synthesis", pp. 255-263 in Bailly & Benoit (Eds) *Talking Machines*, North Holland, 1992.

[8] S. Öhman "Coarticulation in VCV utterances: spectrographic measurements", Journal of the Acoustical Society of America 39, 151-168. 1965.