Acoustical Society of America and Acoustical Society of Japan

Third Joint Meeting

Honolulu, Hawaii, 2-6 December 1996

CHATR: A High-Definition Speech Re-Sequencing System

Nick Campbell
ATR Interpreting Telecommunications Research Laboratories
Hikari-dai 2-2, Seika-cho, Kyoto 619-02 Japan. (nick@itl.atr.co.jp)

This paper describes a method for producing speech synthesis, without the need for signal processing, by using re-sequencing of carefully selected phone-sized segments from a pre-recorded speech corpus. The generated speech faithfully reproduces the voice characteristics and speaking style of the original to create novel utterances.

The process involves creating an index of phones and their prosodic characteristics for each utterance in the corpus. The re-sequencing synthesiser doesn't necessarily produce any sounds; it merely determines an optimal sequence for random-access replay from the original speech to give the best approximation to a desired utterance from the segments available in a given speech corpus. The synthesis method is independent of language or speaker but requires a sufficient source database that represents a balanced sample of the language

To find the optimal sequence of segments for concatenation, the synthesiser selects from amongst candidates in the database using a weighted combination of their acoustic and prosodic features to maximize continuity between segments while at the same time minimising the distance of each from a given prosodic target. Optimal performance is achieved by under-specification of prosody, so that only key points in the utterance have targets and the remainder are considered prosodically neutral. In conjunction with loose selection of units from a continuous-speech corpus, prosodic under-specification maximises the number of candidate segments and uses the redundancy of information in natural speech to reduce or eliminate distortions in the output synthesis.

1 Background

Conventional speech synthesis methods use a limited inventory of phonemes, diphones, or demisyllables as basic speech units. To account for contextual phonetic effects from surrounding segments, Nakajima and Hamada [12, 6] proposed a Context-Oriented-Clustering method to automatically produce an optimal source unit set by using a statistical clustering technique. For Japanese, this resulted in approximately 1500 speech units. At the same time, Sagisaka proposed a scheme for the selection of non-uniform units [13, 15, 14, 11], to be excised at synthesis time from a database of 5000 recorded words. The units were selected

according to a weighted minimisation of contextual spectral distance, acoustic typicality, concatenation cost, and a spectral discontinuity measure [14]. Both the above methods used parametric coding of source units to allow modification after concatenation to warp the unit sequence to the desired intonation.

Parameterisation of the speech waveform can be major source of distortion in the output speech because of a) oversimplified pulse/noise excitation, and b) mismatch between vocal tract spectra and source spectra arising from the prosodic modification. Hirokawa [7, 8, 9] proposed a large-database waveform-dictionary approach to concatenative speech synthesis as a solution to this problem, and achieved high-quality speech output by using a very large inventory of source waveforms. He recorded two hours of speech from an adult male to produce a dictionary of 35000 phoneme-length waveform segments identified by duration, average pitch, pitch contour and average energy. They were classified under 35 phonemic labels and selected by matching both prosodic (accent type and position in breath group) and phonetic attributes. The acoustic phonetic segments were determined manually and separated at zero-crossing points closest to the phoneme boundary labels before being sub-categorised according to prosodic criteria. For selection, pitch was weighted more heavily than amplitude or duration because of the relative difficulty of modifying the former. An evaluation function was used to compute for each candidate segment the difference between the desired prosody pattern for synthesis and the waveform prosody characteristics, making use of an experimentally defined balance factor for the weighting between 'active selection', based on the phonetic environment, and 'static selection', derived from the prosody. When no candidate waveforms exceeded a pre-determined selection threshold, parametric synthesis was used to create an appropriate segment.

The present work extends the above methods by automating dictionary construction and thus enabling the use of any arbitrary speech database as a source for synthesis units, and by defining context-specific selection criteria that eliminate the need for signal processing.

2 Prosodic under-specification

It is clear from the above that high quality synthesis depends both on appropriate segmental and prosodic contexts for each source unit. Given an infinitely large source corpus, and an efficient index into it, we could no doubt produce concatenative synthesis that is indistinguishable from human speech, but with a medium-sized corpus, e.g., containing only about thirty-minutes of varied continuous natural speech, we must optimise the selection of candidate segments so that discontinuities between adjacent segments are minimised on concatenation while at the same time selecting only from those that already meet the intonational requirements. However, because the ear is insensitive to small differences in duration, power and pitch, we can often substitute close alternatives to the prosodically ideal segment without recourse to waveform surgery. We can therefore take advantage of the considerable variation in prosodic realisation of different speakers, and of the redundancy of information cues in speech, to provide *adequate cues* only at *key points* in the prosodic contour of a synthesised utterance.

By under-specifying the prosody except at phrase boundaries and accent peaks we open up greater freedom of choice for intermediate candidate segments so that *continuity*

in the prosodic and spectral domains can dominate throughout unmarked portions of the speech. By preferential weighting at accents and boundaries, we ensure that the concatenated speech contains appropriate spectral as well as prosodic cues, such as increased spectral tilt at prominences, creak and natural power drop pre-pausally, and breath intake utterance initially. By including 'accent' and 'boundary' positions as features in the selection targets, and using a five-phone search window to define 'context' ($current \pm 2$), we increase the weighting on these 'prosodic' criteria without recourse to heuristics. The 'smoothing' between segments can then be left to a Viterbi selection process that favours maximally similar candidates (both spectrally and prosodically).

Since the average prediction error for segmental duration is currently reported to be about 25 msec for vowels and 15 msec for consonants (and for power 2.2 and 2.5 dB RMS respectively), then rather than force the waveform segments to values that we (only approximately) predict for a given context, and thereby induce distortion, we consider it preferable to accept a rather loose match between predicted and actual prosody and instead to use that flexibility to minimise discontinuity in the joins between segments. Because segments are concatenated from natural continuous speech sources, key prosodic events are redundantly marked by spectral as well as intonational parameters so the percept of the resulting speech will be more natural and the variation in unmarked locations perhaps overlooked, as in human production.

3 Synthesis as re-sequencing

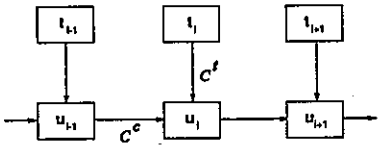
It is customary to consider source units as an integral part of the synthesiser, but by annotating a pre-existing speech corpus with an index for each phon(em)e according its prosodic environment we produce an interchangeable external source. The synthesiser then becomes a retrieval device for random-access re-sequencing that is independent of the source corpus. By this step the synthesis is freed from language-dependencies and from the need to explicitly model speaking style, voice quality, or individual prosodic variation. All are determined as a consequence of the choice of corpus, requiring only a mapping between the transcription labelling and a specification of the target utterance for synthesis. Prosodic details are specified as z-scores, so that selection in terms of speaker norms will naturally constrain all parameters to within the range of each individual corpus.

Chatr synthesis [2, 4] relies on the fact that a speech segment can be uniquely described by the joint specification of its phonemic and prosodic environmental characteristics. The synthesiser performs a retrieval function, first predicting whatever information is needed to complete a specification from an arbitrary level of input and then indicating the database segments that best match the predicted target specifications. The basic requirement for input is a sequence of phone labels, with associated fundamental frequency, amplitudes, and durations for each. If only words are specified in the input, then their component phones will be generated from the lexicon or by rule; if no prosodic specification is given, then a default intonation will be predicted from the information available.

Chatr pre-processing of a new source database has two stages. First, an *analysis stage* that takes as input an arbitrary speech corpus with (at least) an orthographical transcription, and produces a feature vector describing the prosodic and acoustic attributes of each phone in that corpus. Second, a *weight-training stage* that takes as input the

feature vector and a waveform representation, and produces from it a set of weight vectors that describe the contribution of each feature towards predicting the best match to a given target specification.

At synthesis time, the *selection stage* takes as input the feature vectors, the weight vectors, and a specification of the target utterance, to produce an index into the speech corpus for random-access replay to produce the target utterance. For a given target specification, $t_1^n = (t_1, ..., t_n)$, the optimal set of units $u_1^n = (u_1, ..., u_n)$ is selected from the corpus so that the desired prosodic characteristics are realised and the units concatenate together smoothly with minimal distortion. The units are determined by a Viterbi search of the closest n candidates to find the path with the cheapest cost, minimising both distance from target and continuity distortion between segments[3, 10].



Minimising two distance measures

3.1 Two distance measures

The *target cost*, $C^t(u_i, t_i)$, is an estimate of the difference between a database unit u_i and the target t_i it is to represent. The cost is calculated as the weighted sum of the differences between the elements of the target and candidate feature vectors: these differences are the p *target sub-costs*, $C_j^t(t_i, u_i)$ $(j = 1, ..., p)$. The number of features, p, is database-specific, but typically about 25. 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* $C^c(u_{i-1}, u_i)$, is an estimate of the quality of a join between consecutive units (u_{i-1} and u_i), determined by the weighted sum of q *concatenation sub-costs*, $C_j^c(u_{i-1}, u_i)$ $(j = 1, ..., q)$. The sub-costs are calculated for u_{i-1} and u_i from distances of vector-quantised cepstral measures at the point of concatenation, from absolute differences in log power and pitch, and from differences in the estimated R-K voice-source parameters TL, OQ, and GN [5].

The concatenation cost, given weights w_j^c, is calculated as below. 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 aim of the unit selection is to find the best combination of units $\overline{u}_1^n$ from those available in the database that come closest to the specification of the novel utterance for synthesis. Selection of the optimal unit sequence is achieved by minimising the total cost $\overline{u}_1^n = \min_{u_1,...,u_n} C(t_1^n, u_1^n)$. The total cost for a sequence of n units is the sum of the target and concatenation costs and subcosts for each of the features.

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

3.2 Weight training

The weight-training stage of database analysis provides a phone-dependent weight set (different weights, w_j^t, for selecting different phones) by a three-stage process.

The weights for determining the target costs and sub-costs are calculated using the waveforms of the natural speech which are available in the synthesis database. Treating the database as a closed set for training, we

employ a one-held-out algorithm to find the closest equivalents for each phone in turn. By ranking these according to an objective distance measure we can determine the contribution of each of the descriptive features to the selection of an optimal substitute phone. In synthesis, when no such objective target is available, we can use the same weighting of features to select an optimal segment sequence for the synthesis of a novel utterance.

Stage one of the weight training requires describing each phone in the database in terms of the features that can be used for selection. These include phonetic features, such as place and manner of articulation, and prosodic features such as the pitch, duration and energy of the current and neighbouring segments. Stage two of the training holds out each phone in turn as a potential target and ranks the remainder by a time-normalised weighted Euclidean cepstral distance to determine which features contribute most in selecting the best candidate. Stage three uses linear regression analysis to assign weights for each feature according to the importance of each in the various contexts of selection [1, 10].

4 Summary

Chatr is a system for synthesing speech from a large natural corpus that minimises the processing in order to maximise the naturalness of its output speech. Chatr makes no use of parametric signal coding or explicit manipulation of pitch or duration. It requires instead careful indexing and selection from a very large number of source units. The basic unit of synthesis is the phoneme, relying on the source database to contain a sufficient representation of phones in a variety of prosodic contexts. Selection of the appropriate sequence of units from the database is done so that the phones are both maximally close to the desired prosodic target and at the same time have minimal discontinuity when concatenated. In order to meet these two criteria, the selection weightings are assigned on a feature-by-feature basis for each phone class.

Whereas most of the processing is automatic, and a complete new voice as been made in less than a day, from initial recording to final synthesis, there is still a need for manual intervention in the analysis and training stages. Although auto-aligning is highly developed, there are occasions when the labelling can be improved by manual post-processing, and the quality of the synthesised output then increases accordingly. However, as a result of this manual intervention, it is inevitable that errors will appear in the labels, which complicate the later training and require further manual intervention. The brunt of future work will therefore focus on making the entire process as automatic as possible, incorporating more speech recognition technology to improve the labelling, and reducing the subsequent processing required for synthesis.

Another interesting direction for future work concerns prosodic prediction in the text-to-specification stage of the synthesis. There is a circuitous redundancy when features of the database such as prominence, part-of-speech, and proximity to a boundary are used to predict *e.g.,* duration and pitch, which are in turn used to select units from the corpus. The units will be likely to come from contexts that match the original structural specification. However, since the predictions from the prosodic features can never be perfect, it would be more sensible to characterise the required segments directly in terms of the structural context and do without the intermediate numeric predictions. The units se-

lected from appropriate contexts (described according to a given set of features directly) are most likely to have the durations and pitch characteristics that we were trying to predict anyway. This removes yet another component from the synthesis system, producing a faster simple indexing and selection machine.

Future versions of Chatr will be smaller, doing less work, but requiring more source data. Since Chatr is best suited for closed-domain synthesis, these corpora should not be difficult to collect, but by requiring ever more variety in the types of phone represented, the size of the source corpus will also become a significant consideration. The third area for future work might therefore focus on reducing the size of the database, selecting a rich subset instead of using the entire corpus, but since the trend in computer design is still in favour of larger and faster machines, particularly geared for multimedia access, this can take a lower priority for the time being.

Acknowledgements

The author would like to take this opportunity to thank all present and past members of ITL who contributed to Chatr, especially Paul Taylor, Alan Black, and Andrew Hunt.

References

[1] A. W. Black & W. N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis". In *EUROSPEECH '95.*

[2] W. N. Campbell, "Synthesis units for natural English speech". Technical Report SP 91-129, IEICE, 1992.

[3] W. N. Campbell & A. W. Black, "Prosody and the selection of source units for concatenative synthesis". In J. van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors, *Progress in Speech Synthesis.* Springer Verlag, 1996.

[4] W. N. Campbell & A. W. Black, "CHATR: a multilingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, 1996.

[5] W. Ding & W. N. Campbell, "Detection of sentence prominence using voice source parameters", ASA-ASJ '96, this volume.

[6] K. Hakoda, S. Nakajima, & H. Mizuno, "A new Japanese text-to-speech synthesiser based on COC synthesis method", pages 809-812, Proc ICSLP, 1990.

[7] T. Hirokawa, "Speech synthesis using a waveform dictionary", pages 140-143, Proc Eurospeech, 1989.

[8] T. Hirokawa & K. Hakoda, "Segment selection and pitch modification for high quality speech synthesis using waveform segments", 337-340, Proc ICSLP, 1990.

[9] T. Hirokawa, K. Itoh & H. Sato, "High quality speech synthesis based on Wavelet compilation of phoneme segments", 567-570, Proc ICSLP, 1992.

[10] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database". In *ICASSP '96*, 1996.

[11] N. Iwahashi, N. Kaiki, and Y. Sagisaka, "Concatenative speech synthesis by minimum distortion criteria". In *ICASSP '92*, II–65–68, 1992.

[12] S. Nakajima & H. Hamada, "Automatic generation of synthesis units based on context-oriented clustering", S14.2, Proc ICASSP, 1988.

[13] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units". 679-682, Proc ICASSP, 1988.

[14] Y. Sagisaka & N. Iwahashi, "Objective optimisation in algorithms for text-to-speech synthesis", 685-706 in Klein & Paliwal (Eds) Speech Coding and Synthesis, Elsevier, 1995.

[15] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. "ATR ν-talk speech synthesis system". In *Proc. ICSLP*, pages 483–486, 1992.