音声言語情報処理　17－19
（１９９７．７．１９）

Processing a Speech Corpus for Synthesis with Chatr

Nick Campbell

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

Abstract
This paper discusses the use of phonetic and prosodic labelling in corpora for speech synthesis, and argues for a multi-level approach for the description of speech segments. In contrast to traditional phonetic transcriptions of speech, we include prosodic context as a key descriptor of acoustic variance.

Meaningful variations in minimally distinguishable sounds of situated human speech depend both on the context and on the manner of articulation, but can be indexed very precisely by a small number of higher-level features in combination. For synthesis, we use these features in weight-training and unit selection to determine an optimal sequence of segments for concatenation from a speech database. However, we are faced with the paradox that while larger, higher-level units such as the syllable are ideal for defining acoustic variation, smaller sub-phonemic segments are to be preferred for concatenation.

In order to prepare a speech corpus for use in Chatr, the data must be analysed and processed so that all the essential characteristics of sub-phonemic speech sounds are identified in a way that can be extended for prediction of unseen sentences. The paper describes an improvement of the Chatr synthesis system that incorporates syllable-level labelling with sub-phonemic units, and shows that the multi-level non-segmental approach offers several advantages over the earlier phoneme-based system.

Key words •speech synthesis •databases •prosody processing •syllable-labelling •unit selection

CHATR: 音声合成データベース処理について

ニック キャンベル

ATR 音声翻訳通信研究所
〒619-02 京都府相楽郡精華町光台 2-2
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

あらまし　本報告では音声合成データベースに於いて音素ラベリング及び韻律ラベルリングを用いる方法についてと音声セグメントの多層的な記述方法についてを述べる。従来の音素による音声表記とは対照的に、本方式では音響的な違いを記述する重要な要素として韻律的な文脈を取り入れた。

人間が発する音声の最小限の識別可能な音に於いて、意味を有する音の違いは文脈と調音様式に依存する。しかし、少数の高レベルの特徴を組み合わせることにより、高精度にインデックスを付与することが可能である。音声合成で上記の特徴は「重み学習」と「単位選択」の処理に用いられ、音声データベースより結合のための最適な連続セグメントを決定する。しかし、音響的な違いを定義するには音節などのより大きい、高いレベルの単位が理想的であるが、結合には音素より小さい単位である副音素のセグメントの方が適するというパラドックスになる。

CHATRで使用するデータベースの作成にあたっては、初見で文の予測が可能となる方式で、音素より小さい単位のレベルの音声が持つ重要な特徴が全て識別される様、データの分析及び処理を行なう必要がある。本報告では「音節レベルでのラベルリング」と「音素より小さい単位のレベルでの音声単位」を組み合わすことにより、改良されたCHATRの音声合成システムを記述する。また、音声セグメントに縛られない多層的なアプローチが、音素に基づいた従来のシステムより好結果をもたらすことを示す。

キーワード •音声合成　•韻葎処理　•音声データベース　•音節ラベリング　•音声合成単位選択

1 Introduction

A common application of phonetic and prosodic knowledge in speech technology is for 'media-transfer' — converting information from one medium to another for processing by machine, e.g. in text-to-speech synthesis, or speech recognition. Many applications that make use of these technologies are currently being developed to assist in information processing and to enable more natural modes of access for online data systems such as the internet.

However, because these interfaces are intended for use in multi-media machines, the nature of the discourse is quite different from that envisaged by the early developers of text-to-speech synthesis systems, and the task has changed from one of 'reading text aloud' to 'an interaction with the user for the manipulation and use of on-screen or otherwise mutually-available information'. In other words, the nature of the speaking style expected from a machine has changed from 'clear expository' to 'personal and friendly'. This places greater demands on the synthesisers of the future, which will be required to express finer subtleties of meaning and to produce tones of voice that were typically not needed for the simple reading-out of printed text.

1.1 Basic elements of speech

In human-to-human speech, the transfer of lexical information is only a small part of the total meaningful information exchanged, and the way a given utterance is spoken, its rhythms and prosody and its 'tone-of-voice' are of equal importance. In synthesis, we need to manipulate the intonation, the voice, and the speaking style so as to signal to the listener *how* a given utterance is to be interpreted.

In addition, we need also to be able to make use of the non-linguistic sounds common in speech (such as laughs, grunts, hisses, silence, filled pauses, etc) to signal phatic and meta-discoursal cues such as turn changing, (dis-)agreement, understanding, (dis-)approval, etc.

However, because of the complexity of speech sounds (whether linguistic or not), and the richness of the information they can convey, we prefer not to try to create them by rule or by signal processing, but to re-use natural speech segments in concatenative synthesis, so as to take advantage of their fine acoustic variation for signalling the intended interpretation of an utterance.

We use a cost-based search algorithm to select an optimal sequence of waveform segments from an arbitrary speech corpus (external to the synthesiser) to create novel utterances. For this, we have devised a multi-level labelling system that views the waveform as the result of an interwoven sequence of contoid and vocoid segments united at the level of the syllable.

We label the corpus to encode prosodic variation as an inherent characteristic that distinguishes the speech units. In our view, silence is not 'lack of speech', but an integral part of the speech information, which can be potentially carrying as much discoursal information as the lexical units. We therefore treat 'silent' syllables and pauses in the same way as any other syllable, and label their features accordingly.

1.2 Concatenation of speech segments

Various sizes of waveform segment have been proposed as optimal for concatenative speech synthesis; many early systems used the diphone [10] or demisyllable [6]. More recently, Sagisaka's ν-talk [11] showed the potential of the 'non-uniform' unit to capture context-specific acoustic dependencies for modelling allophonic variation by corpus-based unit selection. Its successor Chatr [3] introduced prosodic features for the initial selection, at the level of the phoneme, to eliminate the potentially damaging effects of subsequent signal processing.

The choice of the phoneme-sized speech waveform segment as an optimal unit for concatenation in Chatr was motivated by the theory that phonemes form the smallest 'building blocks' of the speech signal, and supported by the fact that while a contiguous sequence of single phones naturally forms a 'non-uniform' unit, even a small speech corpus can probably be expected to contain at least one example of every phoneme of the language. By selecting phoneme-sized segments from prosodically and phonetically appropriate larger contexts, we are able to concatenate them into longer natural-sounding sequences of speech. If the contexts are equivalent, the joins will be imperceptible and the exact splice-point of lesser importance.

1.3 Phonemic sequences

For common phoneme sequences, frequent in the corpus, longer runs will be selected, maintaining the elisions and assimilations of fluent speech. In the worst case, when no appropriate token for a desired sound can be found, an approximation can usually be built up from contextually or prosodically less-appropriate phone-sized segments to reconstruct the 'missing' token for synthesis. For example, /pyu/ is an infrequent sequence in Japanese, but would be required to produce the English word [puma]; if no /pyu/ is available in the corpus, a /p/ from a /pi/ context (as in [pima] could be joined to a short /y/ as in [yuki] without noticeable damage. Non-native sounds, for example English consonant clusters like /str/ (non-existent in Japanese) can be built up in the same way. Because the infrequent sequence is by definition 'unusual' in the language, some lack of elision or 'hyper-articulation' of the sequence will be more readily tolerated by the listener.

1.4 Syllable-level segmentation

The use of phoneme-sized speech segments can be considered inappropriate for two reasons: first because prosodic events (in whatever language) take place at the level of the *syllable*, which is thus arguably the smallest meaningful unit of any spoken language, and second because *sub-phonemic* segments are physically more suitable units for concatenation. This paper shows how the Chatr principle of segment-based unit selection can be extended to work with speech labelled at the level of the syllable, while selecting waveform segments smaller than the phoneme. The advantage of the former is a considerable reduction in index size, and of the latter a reduction in the number of perceptible discontinuities in the concatenated speech.

The syllable-level analysis of speech posits a multi-tiered interaction between higher-level cognitive aspects of the command chain and lower-level mechanical aspects of the production process. It has been shown that this separation is appropriate for predicting e.g., the rhythms of speech timing [2], where a small number of linguistic and structural parameters are adequate to predict the syllable-level duration, leaving the details of segmental duration to be determined by a process of accommodation into the higher-level syllable framework.

2 Multi-tiered representations

Significant prosodic events such as stress, accentuation, and tonal variation act at the syllable level, with primary effects on the vocalic peak, and only secondary consonantal effects. Öhman [9] and others have described speech as a bi-level co-production process with consonantal gestures superimposed on an underlying vocalic base. This view offers a way of simplifying the phonological inventory but does not take into account the effects of the prosodic environment which determines the way a given articulatory sequence will be actually produced (i.e., its waveform characteristics).

Like Öhman, we view the speech signal as an interrupted vocalic sequence of varying 'colour' (e.g., formant structure), categorizing both the vocoid peaks and the contoid interruptions by the place and manner of their articulation. In this way we can parametrically encode (or index) waveform characteristics for later selection and synthesis, but at the granularity of the syllable, rather than as individual segments.

When labelling at the level of the phoneme, prosodic characteristics were ill-matched (being more relevant to vowels) but by including fields that encode the prominence and boundary characteristics of each syllable [1], we capture the finer articulatory differences that distinguish phonologically similar sequences. This in turn allows us to select syllabic elements directly according to their prosodic features, removing the necessity for prediction of a prosodic contour as an intermediate stage in the determination of appropriate units.

2.1 Syllabification

Traditional phonological theory (eg [7]) favors maximal onset and avoids ambi-syllabicity of segments, but as our goal is to index fine differences in acoustic variation by a small number of higher-level features, we prefer to describe both the onset and the offset characteristics of each sonorant peak in terms of as wide a context of influence as possible. For example, the first vowel in [banana] is likely to be nasalised even though the /n/ is theoretically in the 'following' syllable. Similarly, in fluent speech the /s/ in [last time] is likely to be shorter than that in [last month] even though both syllable and word boundaries 'separate' it from the following /t/ and /m/.

Since we label speech as an alternation of contoid and vocoid centres, rather than positing any absolute segment boundaries, in synthesis we can locate appropriate 'centres' and join them by overlapping. 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, the joins should be imperceptible. Although we thus define two tiers of phonation, each having effects on the other, the prosodic environment has stonger relations with the vocalic tier and so it is the syllabic peaks that form the core of our index.

In Chatr, when selecting a unit for synthesis or performing off-line weight training, the previous and following contexts of each candidate unit are always considered (with their features tabulated). If the unit is phoneme-sized, then inter-vocalic assimilation is blocked and we are unable to model long-range effects. For example, in RP English, the schwa in [th*e* songs] is pronounced differently from that in [the singing] (less fronted) from anticipation of the following vowel, in spite of their both having the same immediate phonemic context. By indexing the waveform segments at the level of the syllable we can capture such interactions and at the same time benefit from a uniform prosodic environment.

2.2 Segment inventories

Because no small corpus can be expected to contain examples of every type of phone in every possible context, we have to index the segments in a way that allows us to find closest equivalents for missing types. For example, in English, the /g/ in [glove] can be imperceptibly replaced with a /d/ which is similar in enough of the significant features. To facilitate such substitution, we avoid unique names (such as 'a', 'i', 'u', or 'p', 't', 'k') for the syllable components, and index them by features instead.

To illustrate with the case of Japanese, a one-hour database of read speech from two short stories published as a cassette book [12], annotated according to the traditional phonemic inventory gives 35 segment types from a total of 30,173 transcribed segments: 5 pure vowels (V), 2 devoiced vowels, 1 semi-vowel, 17 solo consonants (C), 8 geminate consonants (CC), and the nasalised vowel N (and silence). The most common sequence of segments is CV, but CCV, V, VV, VN, CVV, and CVN sequences are also frequently found (131 syllabic combinations). A triphone model for speech recognition on this corpus generates 2940 different models, of which 847 are unique and only 1121 occur more than five times.

2.3 Feature-based encoding

To reduce the number of unique types for feature sharing, we maximise similarities in articulatory feature-space and maximise differences by prosodic characteristics. Acoustically similar doubled vowels and geminate consonants, for example, can be clustered and distinguished by their durations. The nasal N and the palatising semi-vowel, like rhotacisation and 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 vocalic 'carrier' (*V/*), or syllable peak, is well described in low-dimensional space (see the IPA vowel triangle, or F1/F2 formant plots for example) but requires prosodic annotation for a fuller specification. Loudness, duration, fundamental frequency, 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 (*C/*) is also well described by a small number of coordinates in a Cartesian space, dominated by two features: 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 the *V/* and the *C/*; the speech is represented as a sequence of syllables (starting with a silence syllable) such that the onset characteristics of each subsequent syllable will be the *C/* or coda of the previous.

2.4 Sub-phonemic representations

While the higher-level indexing enables wider scope of context and weaker matching of segments in the primary index, it covers relatively larger chunks of the speech waveform. This has the advantage of reducing the size of the index, but requires a mapping from syllable parts to waveform segments. Chatr has shown that phone-sized units are as effective as non-uniform units in concatenative synthesis, but sub-phonemic segments offer greater flexibility for waveform generation if a suitable sequence can be predicted. By using smaller units, selection can be made to 'smooth' across joins that would be disjunct at the phone level when ideal tokens are not available in the corpus.

To provide start and end points in elapsed time for each segment, initial segmentation of the corpus is performed by three-state monophone HMM alignment using a phone sequence generated from the transcription of the utterance[1]. Rather than the single phone mod-

[1] Unlike the data sparcity problem in speech recognition, full use of the whole database can be made for 'training' since the object of the segmentation is to provide an index back into the speech. Re-estimation of pre-trained models on each new database allows a fine fit as long as the orthography matches the spoken sequence.

els used in earlier versions of Chatr, we use smaller 'core' and 'edge' models for finer segmentation. For example, the phone sequence /#+p+y+u+m+a+#/ ([puma]) becomes /# #-p p p-y y y-u u u-m m m-a a a-# #/, labelling both the transition portions and the steady states of the speech stream. From a 7 phoneme sequence, we thus derive a 3 syllable sequence, counting the initial silence, and the two vowels.

The inventory of allowable sub-segment types is #-c, #-0, c, c-c, c-v, v, v-c, or v-0, where #='silence', v='vowel', c='consonant', and 0='null'. This last state is required for the case of v-v sequences, where two syllables are not separated by an intervening consonant, and at utterance endings. As noted above, both '#' and 'v' are treated equivalently as syllabic peaks.

3 Feature-based weight-training

There is little reason to use prosodic targets in the selection of contoid sequences, so for unit selection, once prosodically appropriate vocalic segments have been determined, the join cost [4] alone will suffice for smoothing the sub-phones between the syllable peaks. With syllable-level labelling we relegate the consonantal tier to secondary importance and perform weight training only for the core vocalic portions of each syllable.

The weight training in Chatr [5] uses a 'one-held-out' method to learn the contributions of each feature to ranked waveform distance measures by linear regression. The previous categorical labelling of units according to phonemic type required a linguist to decide which phonemes were phonologically 'similar' before an objective measure could used to decide the selection weights for each set of phonetic/prosodic contexts. In contrast, by using only the vocalic portions of the waveform, labelled not by vowel names, but by the feature description for the syllable as a whole, we can train weights that learn the effects of both prosodic and consonantal influences simultaneously, and can thereby select syllable peaks from optimal prosodic and consonantal environments.

Since two previous and two following syllable feature entries are given as wider context in both training and unit selection, effects for the position of a segment relative to prosodic boundaries and within a prosodic phrase are modeled without the need for special coding. Sub-phonemic fields in the syllable index entry record times for the cv-v and v-vc transition portions and the following contoid centres, allowing appropriate selection of the intervocalic portions to be concatenated after overlapping.

3.1 Unit selection for synthesis

By labelling a corpus at the finer, sub-phonemic level, we have more choice of units for concatenation. To select an appropriate sequence of units when given novel text for synthesis, we first produce a phone sequence using the dictionary, and then convert again from phones to features to obtain a syllable-level specification (including prosody) of the target utterance. If the desired syllable $(V\prime, C\prime)$ combination cannot be found in the source corpus, then a search is performed to select individual vocoid and contoid segments from a 'similar' environment by minimising the feature distances between candidate and target units. Feature distance metrics are determined per corpus in a manner similar to weight training by calculating waveform distances (or their cepstral equivalents) for each setting in the feature bundle.

4 Discussion

Chatr is a speech synthesis system that relies on intelligent data to produce natural-sounding synthetic speech. There is very little intelligence encoded in the synthesiser itself, which serves primarily as an indexing device to identify a suitable sequence of speech waveform segments that will concatenate to form a novel utterance using the voice and speaking style of the speaker of the source corpus.

With developments in intelligent text-processing, we find different needs for speech synthesis, and can foresee applications that require the machine to interact with a human user in a way that uses 'tone-of-voice' and non-speech sounds as much as lexical and semantic information in a discourse. Since the quality of the articulation depends as much on prosodic context as on phonological type, we include ToBI-like [1] stress/accent, tonal-height, and break-index features on each syllable to encode its utterance-level context. Because silences are treated in the same way as regular syllables, their characteristics (such as breaths, lip-smacks, inhalations, laughs, etc) can be coded in just the same way as other speech-related sounds.

By labelling prosodic characteristics directly in the database, we are not only able to select speech segments that best match the target utterance and express finer details of meaning through appropriate voice quality and intonation, but also to remove a large portion of the processing necessary for segment selection. Chatr made signal post-processing redundant by selecting prosodically appropriate segments, but still required prediction of segment durations, power, and fundamental frequency in order to have a target by which to select the phoneme segments. Now instead of using rules trained from the ToBI labelling of the corpus to predict its prosodic characteristics, we can directly index segments according to that labelling, rendering the prosodic prediction phase also redundant. By including the prosodic context in the unit labels, we ensure that the durations, pitch, and power will be appropriate by default.

5 Conclusion

This paper has described the processing steps required for preparing a new voice in the ATR CHATR speech synthesis system, which uses re-sequencing of speech segments from a large representative speech corpus, without requiring signal processing for subsequent modification of the sounds, so that high-definition voice-quality can be achieved for synthesizing utterances with appropriate prosody.

HMM segmentation is performed as before but at the finer level of the sub-phonemic unit, while information related to the wider context of each segment is encoded in a syllable-level feature vector. The use of feature-based identification of the speech segments allows a non-exact match to be made, giving more flexibility in the selection of units for synthesis.

Evaluation of the system is a problem that still remains to be solved. While perceptual scoring is perhaps the most informative method, we are still looking for an objective measure that will reflect the average listener's opinion. Because of the generally very high quality of the synthesised speech, the occasional mismatch in segment selection can have a disproportionately large perceptual effect.

Early experiments with Chatr used special phoneme-balanced speech corpora to ensure optimal coverage of the sound sequences of each language, but because of the high definition of the synthesised speech, stress was frequently apparent in the resulting voices. Later developments showed that naturally-occurring speech samples, such as from 'cassette books', included more expressive intonation (though not as variable as that found in lively spontaneous speech), and fewer 'tongue-twisting' phone-sequences that while ensuring 'balance' also tended to induce stress as the reader struggled to produce these infrequent coarticulations.

While the improvements in corpus design brought similar improvements in voice quality, the problems of labelling at the level of the phoneme meant that much of the context encoding was inefficient. The improvements detailed in this paper have enabled a uniform higher-level context encoding that allows phrasal position and tonal characteristics to be encoded as a feature using sub-phonemic waveform labels.

表 1: Example features used in the current implementation (both binary and n-ary values are used).

vocoid features	n-values
place	3
height	5
rounding	1
nasalisation	1
lateralisation	1
rhotacisation	1
velarisation	1
doubling	1
voicing	1
tone	5
prominence	3
contoid features	n-values
place	7
strength	5
doubling	1
sibilance	1
voicing	1
break index	5

参考文献

[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 S. D. Isard, "Segment durations in a syllable frame", Journal of Phonetics 19, 47 1991.

[3] W. N. Campbell: "Synthesis Units for Natural English Speech", Transactions of the Institute of Electronics, Information and Communication Engineers, SP 91-129, pp 55 - 62.1992.

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

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

[6] O. Fujimura, M. Macchi, and J. B. Lovins ,"Demisyllables and affixes for speech synthesis", Proc 9th Intl Conf on Acoustics, Madrid, 1977.

[7] R. Lass, *Phonology*, CUP, 1984.

[8] J. Local "Modelling assimilation in non-segmental rule-free synthesis", Labphon 2, 190-224 CUP, 1992.

[9] S. Öhman "Coarticulation in VCV utterances: spectrographic measurements", JASA 39, 151-168. 1965.

[10] J. P. Olive, *Acoustics of American English Speech*, Springer-Verlag, 1993.

[11] K. Takeda, K. Abe, & Y. Sagisaka "On the basic scheme and algorithms in non-uniform-unit speech synthesis", pp. 109-112 in Bailly & Benoit (Eds) *Talking Machines*, North Holland, 1992.

[12] Kuroyanagi Tetsuko reading Mukoda Kuniko's *Funa, Usotsugi tamago*, Shinchosa Cassette Book, ISBN4-10-820116-7, 1989.