社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# 音声合成 CHATR のしくみ

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
http://www.itl.atr.co.jp/chatr

あらまし
CHATR は音声コーパスを用いて音声合成を生成する手法である。本手法は信号処理を施すことなく、音声波形に音響的・韻律的影響を付与する「ゲシュタルト」ラベリングによって適切な音声セグメントを選択する。CHATR は音声コーパスに情報を付与することにより、モデル依存ではなく、自然発話データから直接情報を得る。また、この手法により基本アルゴリズムを変えずに、異なる話者や異なる言語に適用する汎用的な音声合成が実現可能となった。本報告では音声コーパスを 7 段階の処理（音声収録、ラベリングや分析、圧縮や情報符号化、自動学習、韻律予測、単位選択、波形接続）によって連続発話音声データから合成音声を生成する方式を紹介する。

キーワード • 音声合成 • 大規模音声コーパス • 発話様式 • 波形接続 • 自然な発話

# Stages of processing in CHATR speech synthesis

*Nick Campbell*

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

Abstract
CHATR is a corpus-based method for producing speech synthesis, without signal processing, by selecting appropriate speech segments according to a Gestalt labeling which annotates prosodic as well as phonemic influences on the speech waveform. From an engineering point-of-view, the synthesiser is minimal, little more than an indexing device, but the labeling of speech variation in the natural data, rather than modeling it in the synthesiser, has enabled a generic approach to synthesis which easily adapts to new languages and to new speakers with little change to the basic algorithms. This paper describes seven stages of CHATR process-sing of a speech corpus for concatenative synthesis. They include recording, analyzing, encoding, training, predicting, selecting, and finally synthesising, or recreating novel speech using the voice of the corpus speaker according to parameters learnt during the analysis of the corpus.

Key words • speech synthesis • corpus-based • spontaneous speech • concatenation • multilingual

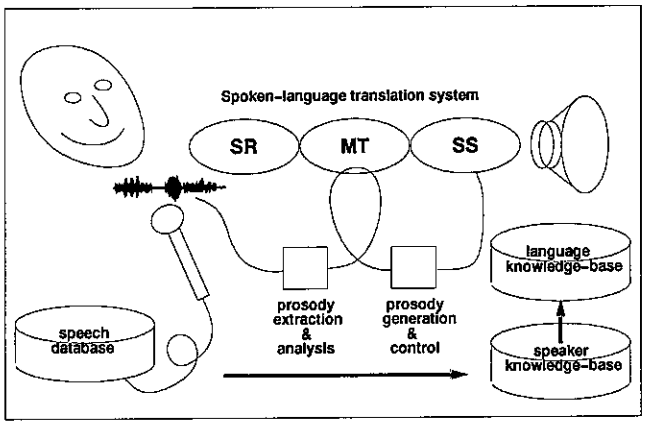


Figure 2: CHATR as part of a speech translation system. (SR: speech recognition, MT: machine translation, SS: speech synthesis). Prosody processing is used for both the extraction and generation of speaking-style information.

# 1 Introduction

The speech synthesis system CHATR [1, 2, 3] was developed as a voice-creation module for an interpreting telecommunications system. It serves as a research workbench for corpus-based speech processing and is being used to test theories of a Gestalt description of speech and prosodic integration.

CHATR is not designed to be a text-to-speech converter, or reading-machine, and works best with annotated text as input in order to produce natural-sounding speech. This annotation information can be obtained from the input speech by prosodic extraction in conjunction with speech recognition and cross-language feature mapping when necessary (see figure 2). CHATR *is* however required to model the characteristics of human conversational speech, and must therefore synthesise not just intelligible speech, but also *human-sounding* speech, including many of the non-speech sounds, such as laughs etc., that are frequent in human spoken communication. Figure 3 illustrates its usage in a translation system, where the text of an utterance is accompanied by information about the speaker's characteristics and about the desired speaking style and meaning.

## 1.1 incorporating prosody

CHATR speech synthesis starts and ends with the speech signal. In order to produce a high-definition rendition of any given input utterance, we need a large corpus of recorded speech samples from which to select small waveform segments for concatenation. The coverage and labelling of these corpora govern the quality of the resulting synthesis [7, 8]. For this reason, prosodic extraction and labelling are integral and essential components of the speech synthesis generation process.

## 1.2 evolution from *ν-talk*

CHATR extends *ν-talk* [9, 10, 11] which in turn extended early concatenative synthesis [5, 6], adopting *ν-talk*'s large-corpus speech unit database approach but incorporating prosodic information as an additional selection criterion, rather than leaving prosodic modification as an after-process. A significant improvement in voice quality is gained by this Gestalt approach to speech segment description, but at the cost of a large increase in the amount of source data required to cover the necessary prosodic variation.

The second main difference from *ν-talk* concerns the type of speech corpus used for source segments. The *ν-talk* system employed a database of 5000 isolated words to provide coverage of all the likely phone combinations expected to occur in the synthesis of Japanese. However, we found that the voice quality resulting from the (probably very boring) task of reading these words resulted in less-than-natural voice quality. Consequently, we now use corpora of more contiguous speech and achieve phonemic and prosodic balance by over-recording and then reducing the database to maximise coverage.

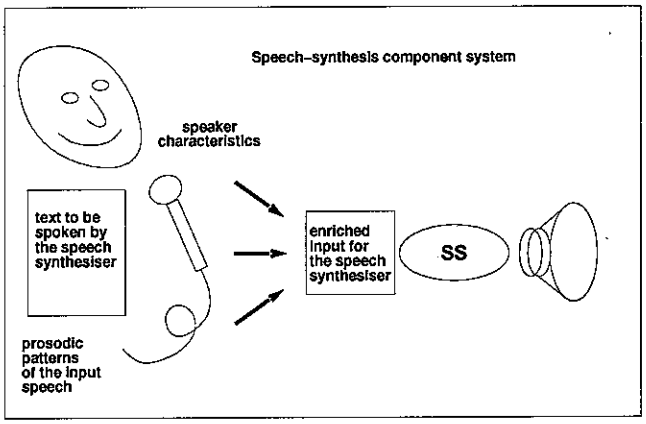


Figure 3: Enriched input for speech synthesis in a speech translation system where information is available about the characteristics of the input speaker and about the speaking style and prosodic characteristics of the input speech.

# 2 Corpus-based processing

Figure 1 summarizes the data flows in CHATR. It shows that processing (illustrated here in the form of pipes) occurs at two main stages: in the initial (off-line) database analysis and encoding stage, to provide index-tables and prosodic knowledge-bases, and in the subsequent (on-line) synthesis stage, for prosody prediction and unit selection. These 'pipes' function as two-way data processing devices, storing encoded representations of embedded knowledge (as neural nets, regression coefficients, or clas-

sification trees) and retrieving closest approximations to target representations by predicting their characteristics and identifying the waveform segments.

In the off-line stage, we start with a speech corpus and its related orthographic transcription and then process the speech to obtain phonemic alignment and prosodic information for each phonemic segment thus defined. This provides the raw material for db_make, a set of programs which produce the index that defines the speech data and allows retrieval of individual segments in an order appropriate for concatenation into a novel utterance.

Because the speech database contains information about not just the speaker, but also the speaking style, the dialect, and the language, we can use the features derived from the text as independent variables in a statistical learning procedure (classification and recursion trees are perhaps the most convenient) and the values derived from the associated speech data as dependent variables for training, in order to learn the characteristics which we will later be required to predict when deciding optimal parameters for synthesis. Thus the speech corpus itself provides a knowledge-base for the synthesis as well as providing a data-base of speech units.

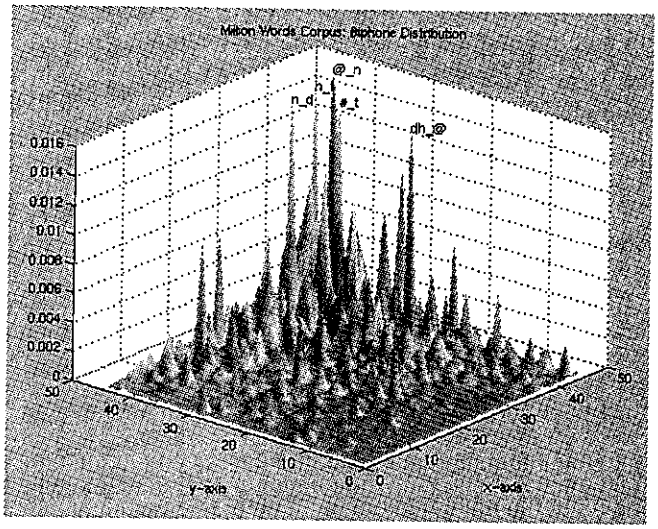


Figure 4: Plot showing bi-phone distributions in a million-word corpus of phonemically-transcribed English text.

# 3 Stages of CHATR processing

This section describes the seven stages of corpus-based speech processing required for CHATR synthesis, summarizes our present understanding of the component technologies, and outlines plans for future work. The stages of processing include recording, analysis, encoding, training, predicting, selecting, and synthesising. The corpus-based approach has enabled a shift of knowledge out of the synthesiser apparatus and into the data, resulting in large, information-rich source databases which are accessed by small index-based search engines. The synthesis process has thus become largely language-independent, but is also speaker- and speaking-style dependent, requiring different databases for each speaker and/or style. Because the processes are entirely data-driven, the generation of new speaker databases requires little external knowledge, and is readily replicable for other speakers and languages.

## 3.1 recording

Since CHATR uses segments of raw unprocessed speech waveform for synthesis units, the quality of the speech database is critical to the quality of the synthesis. It is not important that the recording be of high technical quality, and relatively stationary background noises such as tape hiss may have little effect as long as they concatenate well without noticeable discontinuities. The most important feature for a speech corpus to be used as a source database for CHATR is that it is limited in the range of intonational variation and that it includes a *representative* coverage of the sound combinations for the language and for the desired synthesis task.

Early experiments using phonetically-balanced sets of (5000) words or (503) sentences resulted in speech that was stilted and lacked natural prosodic variation. More recently, after experimenting with spontaneous speech corpora and commercially-recorded cassette-books, we prefer to use contiguous texts such as short stories, dialogues, or contiguous passages that are familiar to the reader and that encourage more expressive voice quality and prosodic variation.

The question of phonemic balance is not one that can be solved mathematically. We can compute the possible phone combinations for a given language and design texts that ensure flat coverage, but this does not take into account the frequencies of occurrence in the daily language and the fact that the more commonly used sequences of sounds can undergo particular articulatory changes through familiarity. Furthermore, explicitly designed corpora can result in a surplus of 'unused' or redundant examples, at the expense of insufficient tokens (and variety) of the commonly occurring combinations which are likely to be more fluently produced. A recent paper [17] reported more than 80 different pronunciations for the word 'and' in a corpus of English, but probably fewer than ten of these would have been predicted by rule. By proportionally representing such sound sequences in the source database, the likelihood of selecting a contextually appropriate variant for synthesis becomes higher.

A forty-minute CHATR database for an English speaker (nes) contains 35,880 triphones, which can be divided into 8,813 groups that differ acoustically from one another. Approximately 45% of all triphones in this database occur only once, or 0.000028 percent of the time. Most prosodically distinct triphones occur very infrequently,

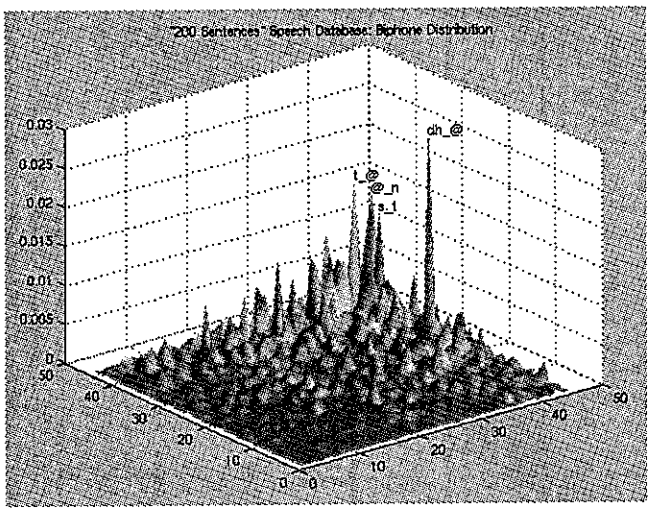


Figure 5: Plot showing bi-phone distributions in a 20-minuet single speaker database of English speech. (the differences are detailed in figure 6

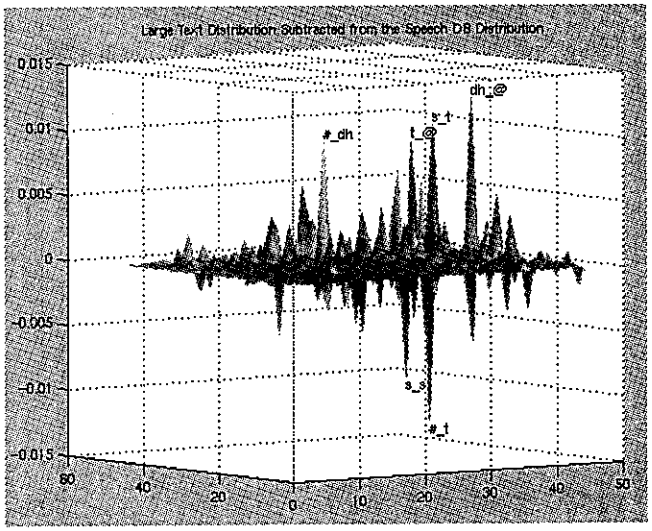


Figure 6: Plot showing difference in bi-phone distributions between a single-speaker (forty-minute) corpus and a million-word corpus of English speech.

and among the 35880 triphones in this speaker database, 34,576 are unique. It is interesting to note that the combination of phonemes "ax", "n" and "d" (the word 'and') is among the most frequent.

However, if we consider biphone coverage instead, the same database contains 35,065 instances, of which only 1,391 are unique. Applying a similar analysis to a much larger corpus, we find that the number of unique biphones appears to saturate at approximately 2,200. The question we must ask is: which biphones occur at what frequencies in speech in general, or in a given task, and what are the features that contribute to their perceptible acoustic variance?

Given a set of 56 unique phonemes 3,136 combinations are theoretically possible, but they do not all occur naturally. Figure 4 shows biphone frequencies for a very large text corpus, and figure 5 shows a similar distribution for one speaker's data. The axes represent phonemes sorted in order of increasing frequency of occurrence in the text. It is of interest to see how flat the graph is over much of its surface; approximately a third of the possible biphones occur with some regularity, and a small number occur very frequently. CHATR should have enough tokens of these regularly occurring sequences that they can be fluently and variably reproduced, preferably including quin-phone context, leaving the unusual combinations to be accounted for by even single phone concatenation, on the fground thateven human speakers ocasioally have difficulty pronouncing the rare combinations. This is the 'non-uniform unit' approach.

Figure 6 shows the differences between a single-speaker corpus and the large-text distributions, which we take to define the language. When collecting speech data, we aim to reproduce the observed frequencies (smoothing the difference) rather than ensure a uniform coverage of the phoneme space.

## 3.2 analysis

When the database has been recorded, and the original text is available, phonemic and prosodic labelling can be performed to produce an index into the speech. Currently we use HMM technology to perform a forced alignment of the speech to the phoneme labels that have been predicted from the text using the preprocessing component of the synthesiser. As an alternative, encouraging results have also been found for DTW-based alignment [16], using the speech waveforms generated by CHATR, from a similar speaker's speech data, to correspond to each sentence in the corpus and then transferring the phone labels after the alignment differences have been minimised.

Similar techniques have also proved successful for prosodic labelling, using multiple transcriptions generated from and constrained by the text, and then selecting the one which best fits the observed prosody of the corpus utterance. We use ToBI transcription [14, 15] to specify the prosody and have found that good approximations can be fitted to the observed contours with this system.

This illustrates the cyclic nature of CHATR processing described in figure 1. The labels used to transcribe the data are generated by the system so the training is self-contained and errors arising from mis-match are minimised. Starting from known facts about the language, we bootstrap the labelling and train similar models for the current database. Because our task is 'closed' we can refine the models iteratively to fit the data.

## 3.3 encoding

The data that has been produced from analysis of the speech corpus now has to be encoded to create an index for the retrieval of appropriate speech segments at synthesis time. The phone labels that were generated by the dictionary are converted into phonetically universal feature bundles, representing place, manner and style of articulation, and non-speech sounds such as laughs, sniffs, noisy breaths, etc., are similarly noted.

The prosodic characteristics extracted as $F_0$ values in Hz, or as durations in milliseconds, RMS power etc., are normalised by z-score transform to produce tables of speaker-specific phone-based means and durations for each feature, leaving the measured values to be expressed as excursion in standard deviation units from a zero mean. By this transform, we can pool data across similar speakers to increase the size of the corpora for training language models, and can use observed prosodic targets for objective testing measures (test_seg, see below) against real speech across different databases.

Database compression is not essential, but we can reduce the number of units in the database by using redundancy measures, and further reduce disk space by waveform compression. MPEG2 encoding offers very high compression ratios and includes perceptual masking of the speech waveform which helps sharpen the objective measures, such as cepstral distance, which are used in selection and database reduction.

Although still experimental, we employ test_seg (i.e., excluding each sentence in turn from the database and taking its phone sequence with values for pitch, power, and duration, etc., as prediction targets for synthesis from the remaining speech data) to compare the closeness of the approximation using objective distance measures. If an utterance in the corpus can be closely replicated using other speech segments from the same corpus, then the speech information it contains can be considered redundant. The closer the replication the more the redundancy, and although a degree of redundancy is desirable in a corpus for synthesis, to ensure sufficient 'natural' variation, if size is a constraint then some pruning may be preferred.

## 3.4 training

Training of the speech corpus for unit selection was described algorithmically in [3], so we can summarise briefly here. We need to overcome the problem of sparcity in any given database, and to do this we must learn which features contribute most towards the selection of an appropriate unit for a given target context. With this information, we can then select from candidates according to their closeness in the feature space, to find an suitable alternative context when the ideal unit is missing from the database.

The procedure is as follows: for each segment in the speech corpus, we use it as a target and list all suitable candidate segments (sorted by closeness according to physical measures of cepstral distance, duration and pitch). The statistical relation between the features used to describe the segment and the ranking of the candidates is learnt by the model and generalised for all phone types in each context type.

Clearly, the choice of features used to describe the phones and their contexts will determine the quality of the subsequent predictions, so the determination of features that best describe the speech is a matter of continuing research.

## 3.5 predicting

For prosodic prediction we have used simple linear regression and neural networks in the past, but our current training predominently uses tree-based methods. Suitable selection and grouping of independent variables is the key to successful prediction, and much work has been devoted to determining optimal feature combinations for first predicting ToBI labels from text input (not needed when the text is annotated) and then for predicting prosodic contours for duration, $F_0$, and power, based on the ToBI sequence.

The Chatr principle is to cut a prosodic image (not separating prosodic and segmental material, since both are produced under the control of the same multi-parameters) into pieces like a puzzle, and then to rebuild the puzzle with pieces from the database. In spite of often selecting units from less than ideal segmental and prosodic contexts, CHATR's generated speech can often be perceived as natural or well-formed. One underlying hypothesis that explains the efficiency of this method is that of the Gestalt functioning of speech. If the perception is processed appropriately, even when some segments of the speech are missing or badly selected, the global contours will still be identified because of 'master segments' which are informationally heavy (since global perception is not linear: information is not shared on equivalent parts). However, if these master pieces are missing, then the perception process fails.

Proximity to a boundary or to a tonal accent will result in significant differences in the manner of phonation of the speech segment. These are better captured by adequate description of the higher-level contributing context than by lower-level fine-phonetic labelling. It is therefore very important that the prosodic environment be well described in order to select a unit sequence for concatenation and synthesis.

## 3.6 selecting

Unit selection requires a compromise. It is rare (and fortunate) for a target sequence to be found easily in the database, so instead we rank candidate units according to two measures of closeness of fit. They must match acoustic and prosodic target trajectories and also ensure smoothness of join between adjacent units.

This selection is made more difficult by the fact that while some discontinuities are more apparent to the ear than to physical measures, the ear can also be insensitive to some other kinds of mismatch. For example, phase differences and some power differences that are obvious to the eye when examining the speech waveform may pass unobserved when listening. A perceptually-based measure is required here, but this too is work in progress [12, 13].

Improvements have been found from the use of locally sensitive selection weights, changing in sensitivity according to their position and context in the utterance, pegging the target cost high in areas of prosodic change and perceptual sensitivity (such as at boundaries and on accents) while freeing it during the intermediate sequences. This has resulted in more freedom for smooth joins and tighter control at perceptually-relevant points than would otherwise have been possible, but requires difficult training.

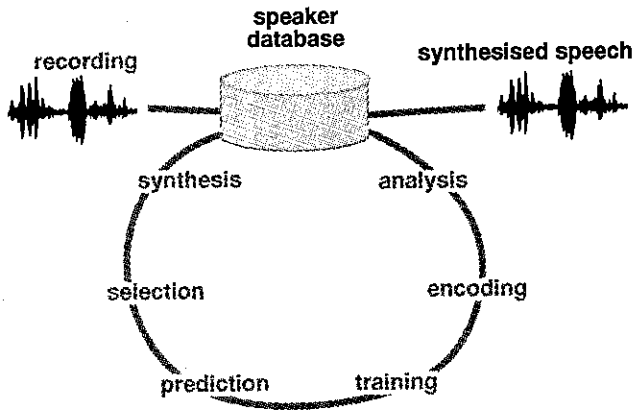


Figure 7: Stages of speech processing in CHATR

### 3.7 synthesising

Waveform concatenation is currently the simplest part of CHATR, as the raw waveform segments pointed to by the index for the selected candidates are simply concatenated and sent to the audio buffer, but it is at this point that post-modification could be performed to improve the prosodic or acoustic characteristics in the case of inadequate units. We have experimented with morphing, ARX resynthesis, PSOLA, and STRAIGHT waveform processing, but all have produced such degradation as to be as disruptive as the discontinuities they aim to repair. Work continues in this area, and as database size increases we may find that signal processing applied to correct the smaller discontinuities will be less disturbing. on the other hand, the smaller discontinuities may then be considered satisfactory without any signal processing.

## 4 Emotional speech

The next major area of speech research for CHATR must concern the expression of emotion in speech. While it is rare for a text-to-speech synthesiser to express emotion, it may be essential for a conversation system to produce appropriate renditions of the input. Too fast a speaking style, and the percept can be one of anger or annoyance. Too much variation in the fundamental frequency, and the percept may be one of over-excitement. Humans are very sensitive to such prosodic and voice-quality changes, and as the synthesis approaches human-like quality, the judgments and reactions of the listener become more critical. Information is parsed as if it were intentional, as it would be with a human speaker, and false impressions can be given.

With this in mind, we have collected three corpora of emotionally marked speech and are studying the acoustic and prosodic correlates of anger, joy, and sadness, so that when the corpora are merged a measure will be found that enables the labelling and appropriate selection of units according to emotional as well as phonemic and prosodic characteristics. Tests have shown that even semantically neutral sentences synthesised from the different databases are easily recognized for the underlying emotion [18, 19].

## 5 Future work

Database processing is essential to efficient concatenative synthesis, and the labeling, analysis, and annotation of speech corpora will continue to be the main core of CHATR research. From a scientific viewpoint, we must focus on determining the features that control variation in speech. in addition to the present segmental and prosodic labeling, there is a need for phonation-style labeling that will help identify speakers' emotional states and speaking styles.

From an engineering viewpoint, the search for a non-disruptive signal processing technique must be continued so that data size can be reduced and prosodic and spectral discontinuities can be repaired. The automatic collection and annotation of speech corpora is also being actively researched, but there is a need for a good perceptually-based measure of closeness between two speech segments if this work is to really succeed. With such a measure, optimization of the weights and pruning of the speech database could be performed automatically and at a higher standard than at present.

There is still a need for human checking of segmental and prosodic labels, and this can be an expensive stage in creation of a new speech database.

# 6 Conclusion

This research initially set out to answer one question: if we remove the conventional constraints on speech synthesisers, such as speed of processing, size of database, etc., then what other limitations will remain to prevent natural-sounding speech synthesis? During the period of our research, the physical limitations on processing speed and memory size have been more than overcome by advances in the hardware.

We have shown that given enough units and with appropriate prosodic specification, we can produce synthetic speech that can be almost indistinguishable from human speech, but we have also found that coverage of the database is the most important limitation to consistent high quality.

No method of signal processing that we have yet tried has been free of damaging effects on the naturalness of the raw speech recording, so if we are to develop CHATR further, then work must be addressed to database design so that optimal coverage of all perceptually significant prosodic and spectral variations can be achieved.

Since there is such a great range of speaking styles and emotions, and since their effects on the signal quality are considerable, and particularly noticeable in high quality recordings (16-bit, 48kHz sampling rate), then a trade-off is apparent. As with speech recognition technology, best quality can be achieved if we constrain the domain. We can achieve flexible synthesis if we accept a loss in definition or a degradation in voice quality, or we can achieve high-quality synthesis if we accept a limitation on the domain of application.

It is my personal belief that general-purpose reading machines are of limited use, and that customized domain-specific synthesis will find many applications in the coming years. If this is the case, then we should work to make data collection and annotation easier and more efficient. In the short-term, this may offer quicker access to high-quality synthesis than parametric or rule-based approaches. A strong scientific understanding of speech must underlie synthesis research, but for the personalization of synthesis systems it may be better to apply that knowledge to the analysis and labelling of corpora rather than to the engineering of speech waveforms, for the amount of information that is present in even the briefest utterance can surely not be modeled by rule, it can only be stored and reproduced.

# Acknowledgments

I would like to take this opportunity to acknowledge Ekaterina Saenko for her work on database design, Kazuyuki Ashimura for his graphics describing CHATR processing flow, and all past and present members of Dept II at ATR-ITL for their contributions to CHATR.

# References

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

[2] W. N. Campbell. "Prosody and the selection of source units for concatenative synthesis". In Proc 2nd ESCA Workshop on Speech Synthesis, Mohonk, N.Y., 1994.

[3] 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).

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

[5] J. P. Olive, (1977), "Rule synthesis of speech from dyadic units", Proc. IEEE-ICASSP77, 568-570.

[6] J. P. Olive, (1980), "A scheme for concatenating units for speech synthesis", Proc. IEEE-ICASSP80, 568-571.

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

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

[9] Sagisaka, Y. (1988), "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. IEEE-ICASSP88, 679-682.

[10] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. "ATR $\nu$-talk speech synthesis system". In *Proc. 1992 Intl. Conf. on Spoken Language Processing*, pages 483–486, Banff, Canada, 1992.

[11] Y. Sagisaka and N. Iwahashi. "Objective optimization in algorithms for text-to-speech synthesis". In *Speech Coding and Synthesis*, W. B. Klein & K. K. Paliwal, Eds., Elsevier Science B. V. 1995.

[12] W. Ding and N. Campbell, "Detection of perceptual discontinuity between phoneme boundaries and its application to unit selection in speech synthesis", Proc. of Meeting of Acoust. Soc. Japan, Yokohama, 1998.

[13] W. Ding and N. Campbell, "Optimising Unit Selection with Voice Source and Formants in the CHATR Speech Synthesis System", Proc. EuroSpeech, Rhodes, Greece, 1997.

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

[15] 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.

[16] W. N. Campbell, "Autolabelling Japanese ToBI" Proc ICSLP-96 (Philadelphia) pp.2399-2402 (1996).

[17] Greenberg, S., "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation", pp 47-56 in Proc. Modelling Pronunciation Variation for Automatic Speech Recognition, Rolduc 1998.

[18] Iida, A., Campbell, N., Iga, S., Higuchi, I, & Yasumura, M., "Acoustic nature and perceptual testing of a corpus of emotional speech", Proc ICSLP-98, forthcoming.

[19] Akemi Iida (Keio University) http://www.sfc.keio.ac.jp/ akeiida/mystudy.html

[20] ATR-ITL CHATR web-page (try it for yourself) http://www.itl.atr.co.jp/chatr/interactive

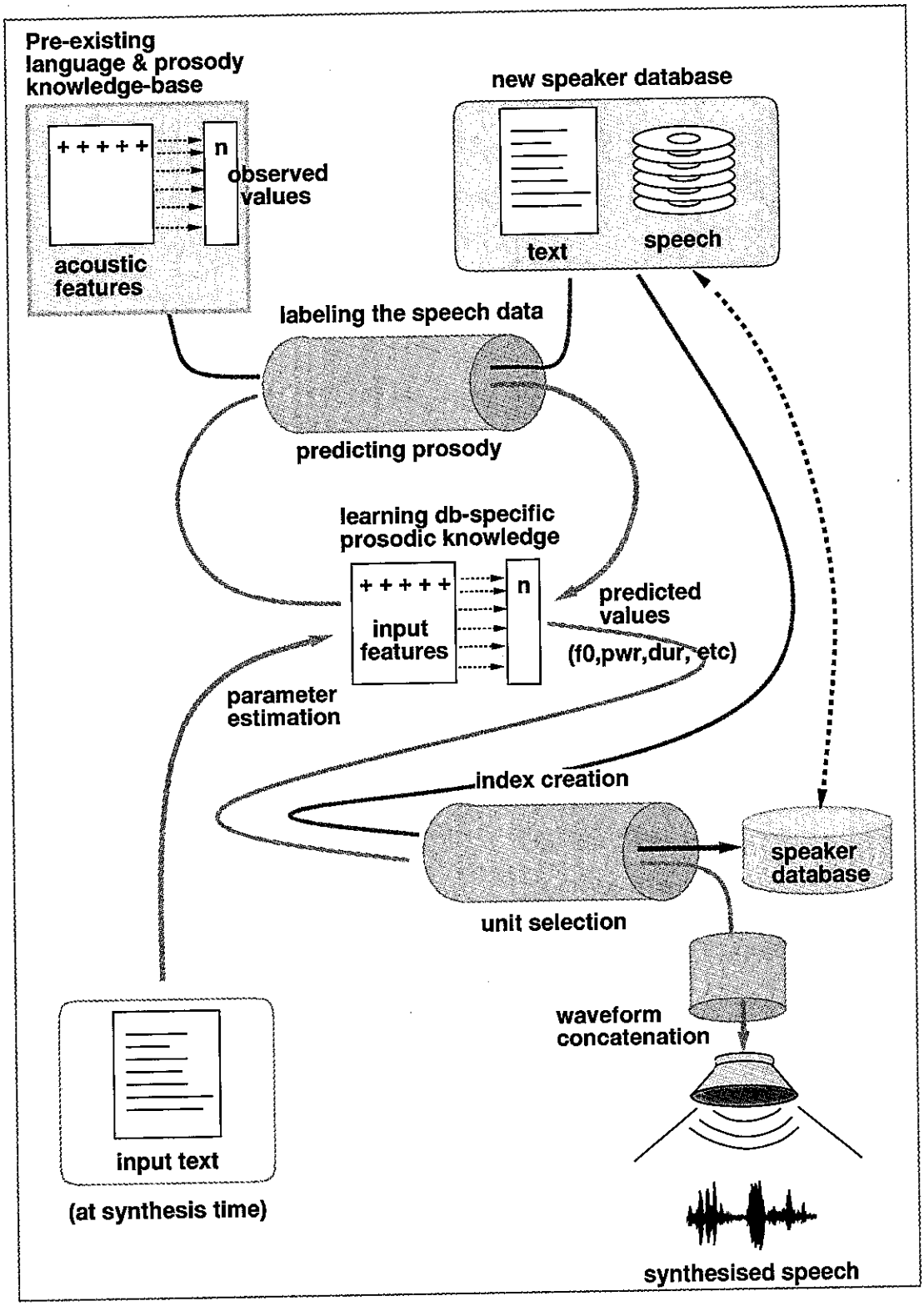


Figure 1: Flow diagram showing CHATR's corpus processing.