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.

Talking Machines for Information Access

Nick Campbell

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

Abstract

Speech synthesis is growing in importance as a method of providing voice-based access to digital information. This paper describes the development of personalisable and friendly-sounding voice-interfaces. In particular, it describes 'DATR', a toolkit for speech database creation for raw-waveform concatenative synthesis, and discusses the features that need to be included when creating a corpus for human-sounding information presentation. In conclusion, it warns of some security and moral issues that arise when using natural-sounding recognisable-voice synthesis techniques, in order that the rights of the indiviual and the responsibility of the source-provider are taken into consideration.

Key words • voice-based interfaces • personalised synthesis • speech databases • features of speech • labelling toolkit • speech copyright

「電子情報と音声合成について」

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

あらまし

声によるコンピュータ情報へのアクセスは、平等なコミュニケーションの方法として有効であろう。情報化社会の進化が生み出したコンピュータのキーボードを自由に操り、自由に情報をアクセスできる人と、コンピュータを使えないがために情報化社会から取り残された人との差を縮めるだろう。社会の国際化や情報化によって、 多言語コミュニケーションはより一般的かつ身近なものになったが、今だ言葉の壁は大きい。本報告で紹介する音声合成とそのデータベース作成技術により、親しみやすい、聞き取りやすい、個人性をもつ合成音声が可能になりました。しかし、この音声合成技術の高品質化に伴い、新たな問題が生まれた。つまり合成された音声があまりに人間の声に近いことによって生じた著作権（声の権利）と、翻訳された結果や声で伝えるニュアンスの相違など内容に対する責任の問題である。

キーワード • 波形接続型音声合成 • 親しみやすい音声合成 • 音声データベース作成 • 音声特徴パラメータ • セグメンテーション • 著作権

1 Introduction

Access to online information through the use of speech synthesis will open up the internet to a wide range of people, many of whom are currently unfamiliar with or hostile towards the use of computers or keyboards but who are often within easy reach of a telephone or radio receiver. Developers of voice-based information-access technology face some interesting challenges in providing services for such listeners, since the current standards of voice quality in telecommunications are much higher than those of synthesised speech. This paper discusses some aspects of the voice-based interface for electronic information and describes recent progress towards a 'friendly' or personalisable rendering device. It also highlights some of the related legal and moral aspects of this emerging technology.

2 Speech synthesis

There has been a shift of paradigm in approaches to speech synthesis. The history of this technology shows an evolution from compute-intensive limited-memory devices towards memory-intensive but light-load systems which make increasing use of natural speech sources for voice-creation. Whereas the majority of speech synthesisers in the nineteen-eighties relied on rule-based approaches, both for the prediction of an appropriate sound sequence and for the production of the speech waveforms, corpus-based developments throughout the nineties have resulted in improved speech quality at the cost of increased memory usage. Immediate improvements were seen when phone-based parametric prediction of waveform spectral and prosodic characterisitcs [e.g., [1] was substituted by diphone-based [2] and non-uniform [3] parameterisation, and similar improvements accompanied the move from parametric to raw-waveform concatenation methods [4, 5, 6].

2.1 Corpus-based speech synthesis

The introduction of large-corpus speech synthesis systems [7, 8, 9, 10] has reduced not only the computational processing load for waveform-generation, but also much of the allophonic and prosodic prediction requirements, shifting the role of synthesis from knowledge-based speech-production to data-based speech indexing and retrieval. This has resulted in naturally-constrained and naturally-varying human-sounding synthetic speech.

The development of large speech corpora for synthesis research throughout the eighties was largely motivated by the need for multiple examples of the acoustic and prosodic variation arising from the different grammatical and phonetic contexts of natural speech. Statistical analysis of these characteristics allowed the prediction of appropriate parameters for intelligible-sounding speech synthesis. A side-effect of this data collection was that the corpora themselves became available for use as a source for synthesis units.

It is a small step from statistical prediction of representative characteristics to direct index-based re-use of segments from the corpora in the synthesis procedure. With large corpora being used as a source of synthesis units, the requirement for statistical prediction of the context-dependent variations is reduced. It becomes possible instead to select a speech unit according to the features of its context rather than having to manipulate its acoustic characteristics in an attempt to replicate the context-specifiic variations. The need for statistical modelling is replaced by the need for representative corpora, and the requirements for extra memory are being met by ever-increasing chip capacities.

Manipulation of prosodic or phonetic information in speech by signal processing is still a non-trivial procedure, and only rarely can it be performed without noticeable degradation of speech quality. On the other hand, it has been shown that the re-use of segments from natural speech can provide very high quality synthesis *iff* appropriate samples can be found in the source database.

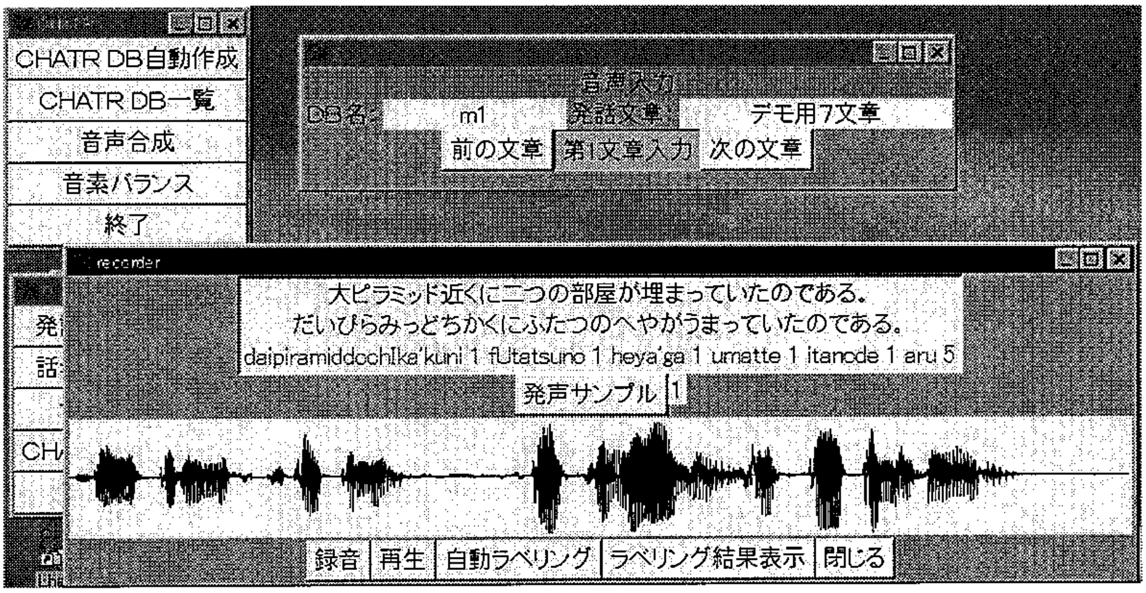
By extension, this process is not just limited to waveform generation, but also to the prediction of the target characteristics; if the database is representative in its coverage, then the need to predict the prosodic and micro-phonetic characteristics of the segment sequence for synthesis is also eliminated. The prediction of context-related speech variation is based on feature-labels describing the corpus segments, but by using direct selection of units according to the same feature-based keys, the natural prosodic and microphonetic variations are constrained to be appropriate by the same laws that ensure their characteristics can be statistically predicted [11].

The logical benefit of such direct feature-based selection is that there is a reduction in the degree of possible error; rather then two levels of potential inaccuracy, first in the prediction and then in the selection stages, direct feature-based selection of waveform segments, with its elimination of the intermediate prosodic characterisation stage, ensures that errors only arise from an inadequacy in the coverage of the corpus.

If the input text for the synthesiser is suitably marked-up with the required features for the expression of its meaning, with phrasal boundary, accent, and focus information, then the processing requirements of the synthesis engine are almost completely eliminated.

2.2 Personalised synthesis

The use of different voices in speech synthesis allows application-specific or customisable 'voice fonts' to be selected in much the same way as screen appearance can be personalised under many computer operating systems. However, with raw-waveform concatenation, the amount of information in the voice is greater than that of para-



DB名: m1 発話文章: デモ用7文章

大ピラミッド近くに二つの部屋が埋まっていたのである。
だいぴらみっどちかくにふたつのへやがうまっていたのである。
daipiramiddochIka'kuni 1 fUtatsuno 1 heya'ga 1 umatte 1 itanode 1 aru 5

Figure 1: DATR's Recording Screen

metric synthesis, and care has to be taken that the 'font' is appropriate to the content of the message. An example of voice-content mismatch can be found at [12], where sadness evident in the voice reading a weather-forecast lends an inappropriate interpretation of the text.

The expression of emotion in the reading of a weather forecast may inapppropriate, but the lack of such expression in the interpretation of a personal message may be more so. Although current synthesis technology is very limited with respect to paralinguistic modelling, the expression of emotion and speaker's attitude is an integral part of the spoken message and future synthesis must include the control of speaking-style and phonation-style in addition to the control of phonetic and prosodic variation if it is to be expressive.

Work is under way on the analysis of the prosodic and acoustic characteristics of emotion in synthesis [13, 14] and on the collection of corpora of emotionally-marked speech [15], but the labelling of paralinguistic characteristics in speech is still very time-consuming as it currently can only be done by human judgement.

3 Speech corpora

The challenges immediately facing high-quality speech synthesis are (a) the design, collection, and processing of sufficient speech corpora for use as source-unit databases, (b) the determination of a sufficient and adequate feature-set for their labelling, and (c) the development of measures for evaluating and maximising the efficiency of the corpus. This section describes components of 'DATR', a suite of database processing tools recently developed at ATR, and discusses elements of database design for concatenative synthesis.

3.1 Tools for database processing

The ATR Database Acoustic-processing Toolkit Resources (DATR) were developed individually for CHATR speech synthesis processing [6], but have recently been merged into a stand-alone piece of software for the recording, labelling, and indexing of speech corpora.

The resources enable a new speaker's voice to be registered for use in the CHATR synthesis system with minimal human intervention or supervision. An untrained user can produce a database of careful speech which has the balance of content required for concatenative synthesis for a given application domain.

Like CHATR's GUI interface [16] DATR is composed of a core set of libraries linked and accessed through a tcl/tk interface, using the Snack [17] speech i/o modules for recording and display of the speech waveform and associated labels. It runs on both UNIX and Windows platforms.

The text corpus is balanced to ensure adequate coveraqe for the task requirements, and the database is recorded interactively by presenting utterance prompts in a sequence that is optimised to fill gaps in the acoustic space according to both domain-specific occurrence-probability statistics and current database content.

The input is prompted at three levels, showing kanji, kana and prosodically annotated phonetic transcriptions for each sentence, with an audio sample also available to indicate the preferred interpretation of the text (see Figure 1). As each utterance is produced, the speaker has a chance to compare it with the audio prompt and re-record, if necessary, before sending it for processing and labelling.

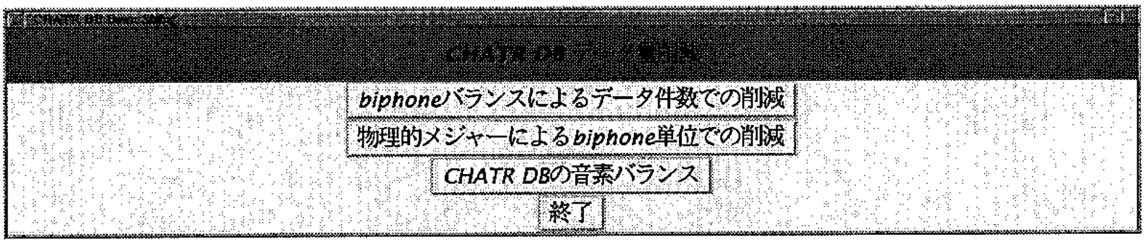

biphoneバランスによるデータ件数での削減
物理的メジャーによるbiphone単位での削減
CHATR DBの音素バランス
終了

Figure 2: Database Reduction Screen from DATR

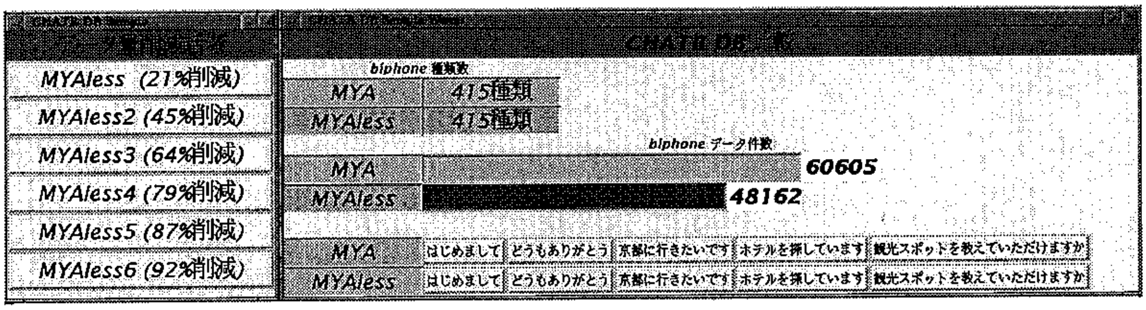


Figure 3: Statistics and sound samples from the reduced database. Listeners can compare the effects of different degrees of pruning by perceptual as well as statistical tests.

3.2 Database design

The optimal design for a CHATR speech database depends on the use to which the synthesis is to be put. For example, a finite domain, such as a number grammar can be calculated to determine the occurrence probabilities of all phonetic and prosodic contexts for any given style of number; the variety of individual numbers may be infinite, but the ways we can say them are predictable. Similarly with weather forecasts, for example, the regions, weather types, and frame sentences form a limited finite set, so their phonetico-prosodic characteristics can be exhaustively computed and phonetic balance guaranteed. Open-ended tasks such as news reading, on the other hand, present a different type of problem; the lexical content cannot be predicted in advance so more general characteristics of the language must be estimated instead.

For predicting the acoustic characteristics of probable sound sequences in news reading, we have processed a large corpus of newspaper texts by first passing them through the 'kan2rom' (kanji-to-romaji with accent and pause information) text pre-processing modules of the synthesiser (see samples below). After calculating the minimum-entropy-based occurrence likelihoods for segments and frequently-occurring segment sequences [18] (see Table 1) we then apply these to determine the information content of the component sounds and prosodic contexts for each sentence in the corpus [19]. This measure of sentence-based acoustic novelty allows us to select sentences for reading in a sequence optimised to provide representative and balanced coverage.

```
    korewa2s-hijoodookoos-hi'daide2joojoo
ga1raine'Ndoni1sakiokurini1naru1kano
oseega1a'ru1kotoo1aki'rakani1-s-h-I-t-a1mo
no5 -60.313850
    do'os-hiwa2se'efu1kIseega1o'oito1s-hij
oowa1toos-h-Ikakara1keeeN1sare1kassee
kaga1sogai1sareruto1kyoochoo1-s-h-I-t-e1i
t-a-5 -71.315048
    yo'toowa2hooaNwa2s-huuiNseejikaika
kUcho'osa1tokubetsu'ide2koNgetsu1hat
sUkani1s-hiNgiiri1s-hi1juuichigatsu1fUt
sUkani1saiketsu1suru1hoos-hiNo1kimete
1ir-u-5 -121.312134
    ka'ikoni1oojita1baai1saiko'yoomo1ar
ie'ruga1chi'NgiNwa2ya'ku1roku'warini1
da'uN1s-h-Ikamo1ichineNgo'toni1koyook
e'eyakuo1koos-hiN1suruto1iu1naiyooda
'-t-t-a-5 -84.344673
```

However, so-called phonemically-balanced sentences can be extremely difficult to read aloud, and since a relaxed voice is usually preferred for synthesis, we do not maximise the compactness of the sentences in order to minimise the number of texts to be read. Instead, we use a larger number of easier-to-read texts that are preferably contiguous. This requires a subsequent reduction in the size of the database to avoid unnecessary duplication of individual segments.

3.3 Database labelling

Acoustic parameter-extraction for segmentation and indexing is performed in the standard way as for speech recognition, and labelling is performed by alignment of

Table 1: Likelihood values for sound sequences in news sentences (the apostrophe represents an accent fall, numbers are prosodic breaks, with '5' indicating the end of a sentence). It is interesting to note that the past-tense marker 'atta' occurs very early in this list, being used more frequently than many single consonants.

-0.0510	o	-1.2460	g	-1.5121	o-5	-1.8874	I
-0.3969	1	-1.2877	d	-1.5130	k-a-5	-1.9390	'-i-5
-0.5067	a	-1.3071	y	-1.5306	s-h	-1.9398	I-t-a-5
-0.5373	'	-1.3222	e-t-a-5	-1.5331	U	-1.9401	e-1-i-r-u-5
-0.6001	i	-1.3249	i-5	-1.5489	r-u	-1.9586	i-N-5
-0.7430	e	-1.3287	'-r-u-5	-1.5509	t-a	-1.9724	N-d-a-5
-0.7697	u	-1.3321	2	-1.5613	'-t-t-a-5	-1.9842	p
-0.7724	N	-1.3645	c	-1.5623	s-o	-1.9880	t-t-a-5
-0.8003	k	-1.3811	t-a-5	-1.5657	s-h-I-t-a-5	-2.0543	s-h-I
-0.8927	6n	-1.3981	s-U-5	-1.6002	o-d-a-5	-2.0947	i-r-u-5
-0.9787	6s	-1.4424	r-u-5	-1.6163	b	-2.1028	t-e-1
-1.0421	t	-1.4464	n-a-i-5	-1.6410	1-s-u-r-u-5	-2.1208	f
-1.1236	k-o	-1.4498	j	-1.7124	N-5	-2.1244	1-i-r-u-5
-1.1270	h	-1.4527	e-5	-1.7188	s-u-5	-2.2078	1-s-h-I-t-a
-1.1710	5	-1.4530	a-r-u-5	-1.7195	a-i-5	-2.3010	1-s-h-I
-1.1881	r	-1.4635	w	-1.7796	a-'-i-5	-2.3177	k-o-n-o
-1.1999	u-5	-1.4913	d-a-5	-1.8131	e-N-5	-2.5498	t-e-1-i
-1.2325	a-t-t-a-5	-1.5078	o-o-5	-1.8169	k-u-5	-2.5553	s-h-I-k-a-'
-1.2332	m	-1.5093	a-5	-1.8444	z	-2.5898	h-I-t

the known phoneme sequence for each utterance. Prosodic feature extraction takes place after segmentation and the details of each utterance are added to the index database until adequate segmental and prosodic coverage is reached. As utterances are added to the corpus, the balance is re-calculated and subsequent utterances are prompted accordingly.

We are still experimenting with segmental and prosodic labelling methods, and modules exist for both hidden-Markov-model-based and DTW-based label alignment. Prosodic feature abstraction uses analysis-by-synthesis techniques based on J-ToBI prediction and verification [20] to produce the high-level description of the segmental contexts.

To the extent that utterances agree with the prescribed transcription, labelling can be performed fully-automatically by these methods. However, since we cannot be guaranteed that the speaker has verified each utterance before sending it for labelling, a post-processing module has been added to detect and flag segments that differ by more than a pre-determined threshold from the samples provided (see Figures 2-5 and next section).

3.4 Database reduction

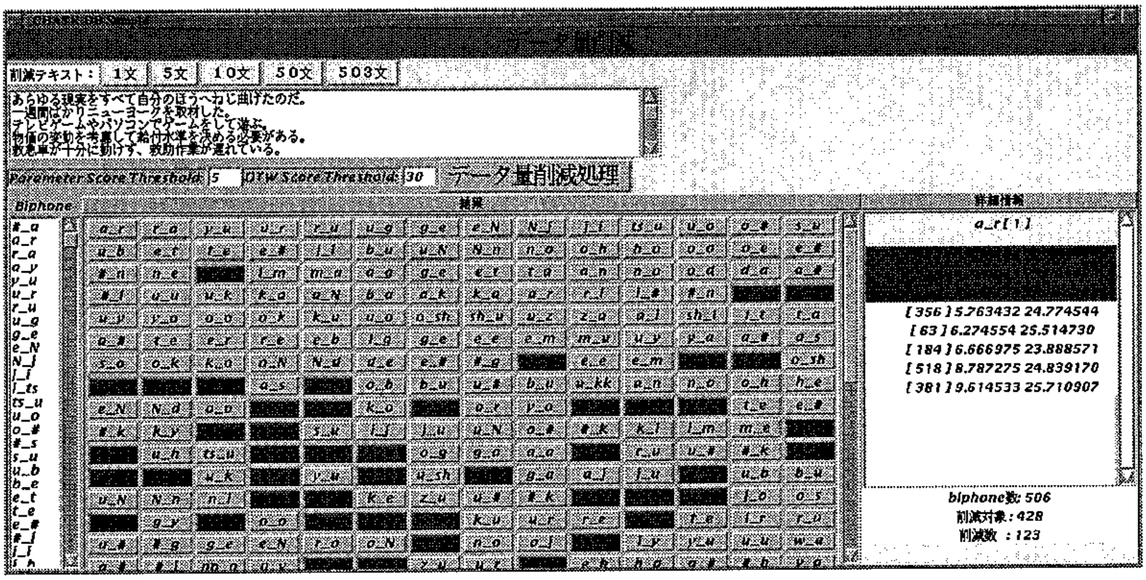
If the speech database contains two segments that are identical (or perceptually equivalent) then only one token need be retained for synthesis, and duplicates should be pruned out for reasons of both elegance and efficiency.

Previous work [21] has shown that objective acoustic measures of the distance between synthesised speech and its naturally-spoken original can correlate well with perceptual evaluations of the synthesis quality. The bispectrum [21] provides a good measure of distance between two signals that are phonemically equivalent. We attribute the efficiency of this measure to the fact that it incorporates phase information, which is particularly important for concatenative synthesis techniques.

In conjunction with the acoustic measure of similarity, we also employ a weighted measure of 4 prosodic features (f0, f0-slope, duration, and power) to measure the closeness between a given pair of phonemically equivalent database segments. Thresholds for the combination of these acoustic and prosodic measures have been determined heuristically.

In DATR, the same measures are used both for evaluating and maximising the efficiency of the corpus. By comparing each segment to its sample original, we can detect mis-labellings and reject suspect segments from inclusion in the index, as well as removing duplicates.

The speech database is pruned by excluding all phone-sized segments that are within a given threshold of distance from another similar segment in terms of both left and right biphone contexts. By adjusting the pruning thresholds, we can determine the efficiency of the resulting speech database along a size-quality continuum. Smaller distance thresholds will produce a larger but more finely-graded corpus, while relaxed thresholds will further reduce the size of the corpus, though possibly at the expense of resulting synthesis quality.



削減テキスト： 1文 5文 10文 50文 503文

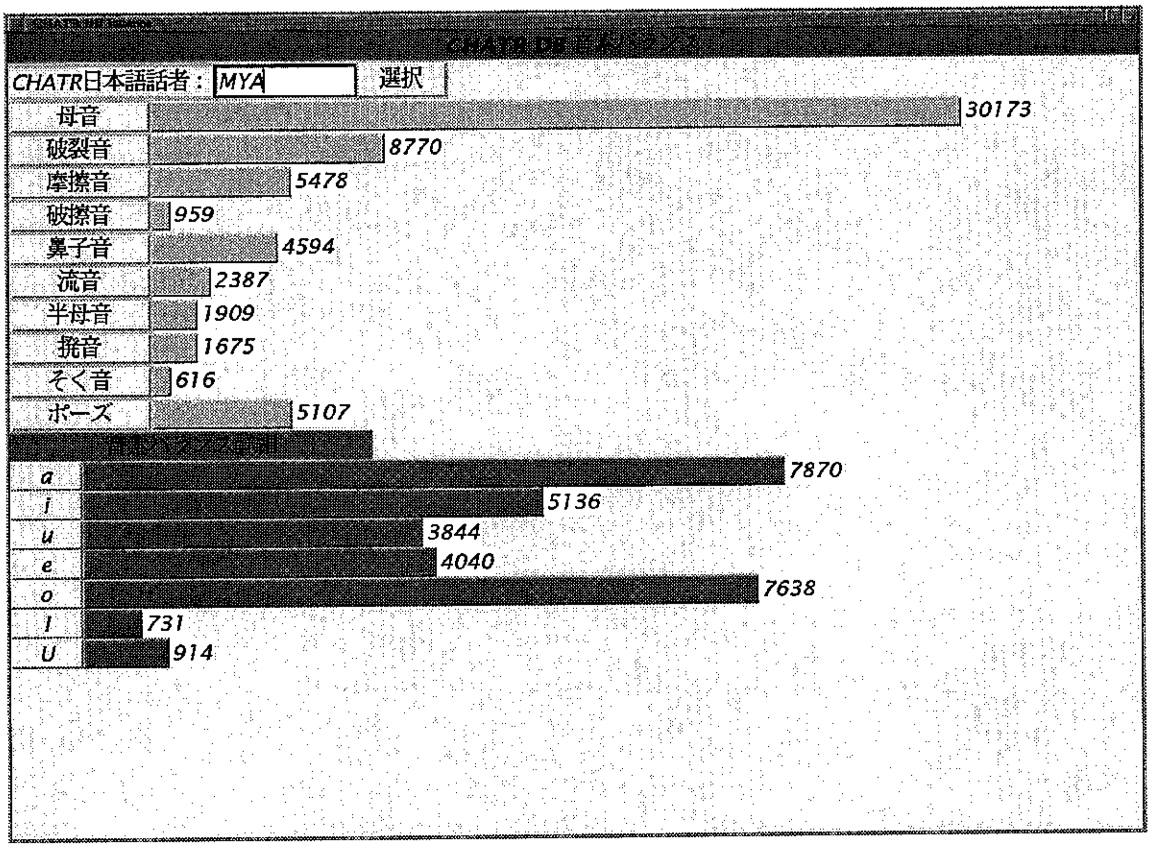
あらゆる現実をすべて自分のほうへねじ曲げたのだ。
一週間ばかりニューヨークを取材した。
テレビゲームやパソコンでゲームをして遊ぶ。
物価の変動を考慮して給付水準を決める必要がある。
救急車が十分に動けず、救助作業が遅れている。

Parameter Score Threshold: 5
DTW Score Threshold: 30

a_r[1]
[356] 5.763432 24.774544
[63] 6.274554 25.514730
[184] 6.666975 23.888571
[518] 8.787275 24.839170
[381] 9.614533 25.710907

biphone数: 506
削減対象 : 428
削減数 : 123

Figure 4: Deleted segments for a given set of thresholds. Darker segments have nbeen flagged for exclusion because they are within a threshold of similarity to other segments already present in the corpus



CHATR日本語話者 : MYA

Figure 5: Segment availability statistics after pruning (counts show only segmental types, not the amount of prosodic variation that they include).

4 Rights & Responsibilities

The use of recognisable and identifiable voices in speech synthesis allows personalisation of information services, but carries with it a potential for abuse. In the interests of the database providers, some issues of security need to addressed before its widespread application. In the meantime, controlled use, i.e., the distribution of pre-synthesised speech samples rather than whole speech databases, is to be preferred.

In the foreseeable future, this technology may enable the replacement of personality voices in broadcasting and audio service provision. There is currently no copyright on voice per se, and the synthesis makes use of small enough chunks of recognisable voice to avoid infringement of any performance rights. However, we should ensure adequate compensation for voice providers, and before recording, speakers need to be informed of potential uses of their speech data.

Because the synthesised voice is recognisable, and as that of a known human speaker, the listener may accept the synthesised information with increased confidence, not necessarily being aware that the source is mechanically generated and therefore prone to computer or other processing errors. Similarly, because the originator of the voice can be identified, the content of the synthesised texts should be monitored to prevent embarrassment or abuse. Providers of voice-based information services should be made aware of their responsibilities and, if possible, laws should be enforced to ensure protection.

5 Conclusion

This paper has described some tools for use in the creation of corpora for raw-waveform concatenative speech synthesis and presented an overview of the current technology. It has stressed the need for balanced and representative source-unit databases for synthesis, and reflected on the paradigm shift wherein the size of the corpus greatly increases computing memory requirements while at the same time reducing the computing load.

It is reassuring to note that this trade-off in the evolution of speech synthesis technology has brought it closer to the characteristics of the human brain, in which the memory-capacity is enormous while the processing speed is extremely slow; the clock-rate of a neural synapse is several centiseconds, while that of a digital computer is measured in nanoseconds.

The basic requirements for human-sounding Japanese speech synthesis are currently met by a J-ToBI-annotated phonemic labelling of read speech but, with increases in database size, we will soon require voice-quality, emotion, and speaking-style labelling as well. Research is currently being carried out on the identification of acoustic characteristics which will allow the automatic labelling of these features for future synthesis systems.

References

[1] Allen, J., Hunnicutt, M. S. & Klatt, D.H. (1987), "From text to speech. The MITalk system", Cambridge University Press, Cambridge UK, 216 pages.

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

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

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

[5] W.N.Campbell, "Labelling an English speech database for prosody control", 1-P-8, Proc ASJ, Spring, 1992.

[6] www.itl.atr.co.jp/chatr

[7] W.N.Campbell, A.W.Black, "CHATR: 自然音声波形接続型任意音声合成システム", 信学技報,SP96-7 1996.

[8] Beutnagel, M., Mohri, M., Riley, M., 'Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis", S4.OR2.1, pp607-610, Proc Eurospeech'99.

[9] F.C.Chou, C.Y.Tseng, & L.S.Lee, "Automatic corpus processing for Mandarin speech synthesis", pp 141-144 in Proc Oriental COCOSDA'99.

[10] Mixdorff, H., Mehnert, D. , "Exploring the Naturalness of Several German High-Quality-Text-to-Speech Systems" pp.1859-1863, Proc Eurospeech'99.

[11] Campbell, W. N., "Processing a speech corpus for CHATR synthesis", Proc ICSP-97, Korea.

[12] www.itl.atr.co.jp/chatr/j_tour/fkt_tenki.html

[13] Cahn, Janet E. The Generation of Affect in Synthesized Speech. In Journal of the American Voice I/O Society, Volume 8. July, 1990. Pages 1-19.

[14] 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.

[15] Emotion and Speech, forthcoming ISCA ETRW, Belfast 2000.

[16] Desirazu & Campbell, "An Extensible Scripting Interface for CHATR", 日本音響学会平成11年度秋季研究発表会 日本音響学会講演論文集 pp309-310 1999

[17] Snack http://www.speech.kth.se/snack/

[18] S. Deligne, F. Yvon, and F. Bimbot. Introducing statistical dependencies and structural constraints in variable-length sequence models. In *Grammatical Inference : Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence 1147, pages 156–167. Springer, 1996.

[19] Campbell, N., & Saenko, E., "Factors to Consider in the Design of an Optimal Speech Corpus for Concatenative Speech Synthesis". Proc ASJ, Mar 1999.

[20] ニック キャンベル「韻律解釈における基本単位」 in 音声と文法II、音声文法研究会、(Spoken Language Working Group) くろしお出版 1999

[21] Chen, J. D., & Campbell, W. N., "Speech Synthesis Evaluation by Objective Distance Measures", in SP-99-xxx, Tech Rept of the IEICE, May 1999.

Talking Machines for Information Access

Nick Campbell

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

Abstract

Speech synthesis is growing in importance as a method of providing voice-based access to digital information. This paper describes the development of personalisable and friendly-sounding voice-interfaces. In particular, it describes 'DATR', a toolkit for speech database creation for raw-waveform concatenative synthesis, and discusses the features that need to be included when creating a corpus for human-sounding information presentation. In conclusion, it warns of some security and moral issues that arise when using natural-sounding recognisable-voice synthesis techniques, in order that the rights of the indiviual and the responsibility of the source-provider are taken into consideration.

Key words • voice-based interfaces • personalised synthesis • speech databases • features of speech • labelling toolkit • speech copyright

「電子情報と音声合成について」

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

あらまし

声によるコンピュータ情報へのアクセスは、平等なコミュニケーションの方法として有効であろう。情報化社会の進化が生み出したコンピュータのキーボードを自由に操り、自由に情報をアクセスできる人と、コンピュータを使えないがために情報化社会から取り残された人との差を縮めるだろう。社会の国際化や情報化によって、 多言語コミュニケーションはより一般的かつ身近なものになったが、今だ言葉の壁は大きい。本報告で紹介する音声合成とそのデータベース作成技術により、親しみやすい、聞き取りやすい、個人性をもつ合成音声が可能になりました。しかし、この音声合成技術の高品質化に伴い、新たな問題が生まれた。つまり合成された音声があまりに人間の声に近いことによって生じた著作権（声の権利）と、翻訳された結果や声で伝えるニュアンスの相違など内容に対する責任の問題である。

キーワード • 波形接続型音声合成 • 親しみやすい音声合成 • 音声データベース作成 • 音声特徴パラメータ • セグメンテーション • 著作権