音声言語情報処理 15－1
（1997. 2. 7）

A step in the direction of synthesising natural-sounding speech*

Nick Campbell
ATR Interpreting Telecommunications Research Laboratories
Kyoto, 619-02, Japan. e-mail: nick@itl.atr.co.jp, http://www.itl.atr.co.jp/chatr

This paper describes speech as a three-part process, portraying meaning through the intersection of text and voice. We distinguish the language-specific aspects of a spoken utterance from the speaker-specific characteristics of the voice, and show how these can be integrated, even for multi-lingual utterances, using a basic synthesis engine that remains flexible and free of language or speaker dependencies.

We present a method for labelling a corpus of speech so that it can be re-cycled to produce novel utterances reproducing the speaking style and voice characteristics of the recorded speaker.

In addition, the paper discusses the use of non-speech sounds that are common in spoken conversation and proposes that these may be an essential part of the perceived naturalness of speech. It suggests that future interfaces using speech synthesis may benefit from the controlled use of such non-speech noises as laughter and audible breaths to ease the interaction with a human user.

1 Introduction

Speech is the transmission of meaning through an oral medium; it differs from writing primarily in that it signals a particular interpretation of the syntactic and semantic content of the text by means of prosodic and voice-quality modulation. Because it is a one-dimensional medium, a function of time, the formal or syntactic structure of speech tends to be simpler than that of written texts, and information signalling the intended interpretation tends to be carried more by the vocal component.

Speech synthesis attempts to simulate the human voice in order to present meaningful information to a human listener directly from a computer. Early work in this field focussed on reproducing the speech sounds ('phonemes') by modelling their waveform characteristics, and later developments concerned methods of predicting intonation and modifying the waveforms to represent the prosodic characteristics of the sounds in different intonational contexts.

Many successful synthesis systems have been presented in the literature [2, 8, 9, 10, 16] and are now being commercially developed but although they are often highly intelligible, and prosodically natural, their output is usually mechanical-sounding and lacks the meaningful variety of human speech [11, 5]. I claim that this is a consequence of focusing research on *intelligibility of the text* rather than on *naturalness of the voice*. By putting emphasis on the written component, we have achieved workable synthesis systems but are only able to convey half of the meaning.

By reversing the priorities, and putting 'naturalness' first instead of 'intelligibility', we have been able to produce a multi-speaker, multi-language synthesis system that generates voice quality that, if the prosody were more fluent, would pass the Turing-test for human-realism. To do this,

*自然音声の合成への一歩
ATR 音声翻訳通信研究所・ニック キャンベル

we have adopted a concatenative approach, that minimises signal processing and recycles segments by random-access from a large corpus of representative recordings. There is a trade-off between natural voice-quality and prosodic flexibility, but this decreases as more source data becomes available. The cost of this approach is therefore an increase in the amount of storage memory required but, as computer memory is constantly becoming cheaper and more available, we feel this is a small price to pay.

The system employs a basic 'core' component to encode the meaning, which consists simply of a relational 'utterance' structure and a set of i/o routines to interface that structure with a library of external modules. These interchangeable external components at various times write the relevant information into parts of the structure, display the relations between them, and produce or modify speech accordingly from them. For more details of the technical design of the Chatr[1] synthesiser, see [7, 8].

Although much research has been put into speech synthesis as a technology, surprisingly little has been done on the types of speech that it will be required to synthesise. It used to be referred to as a 'reading machine' [3] but in practice, the needs for reading written tests aloud are very limited; it now appears much more likely that the synthesiser will be used as a 'talking machine' [2], taking input from less structured and more informal texts such as e-mail and web pages, and having to interact with a human user in a more conversational style that will put higher emphasis on interpersonal aspects of the speech, such as tone-of-voice and speaking style.

In the ATR system, our input is primarily from a spoken translation system, and so we have access to much more information concerning the intended meaning than can be derived from an analysis of the raw text alone. However, to be more generally useful, the synthesiser must be able to accept input from a range of sources, and to adequately portray the intended meaning of an utterance when the input is from text alone. Recent trends in speech recognition have shown the need to process a more spontaneous style of input, rather than read speech; similarly, speech synthesis must now be adapted to cope with the needs of interpersonal communication as well as to the presentation of formally structured information.

In this paper we describe the synthesis process from the point-of-view of information flow, focusing on aspects of the speech that are voice-dependent rather than text-dependent, and discuss the philosophy behind some of the design decisions.

2 'Meaning' – the core

We define the basic unit of speech to be a single 'utterance'. In written text, this would best correspond to a sentence. An utterance is a carrier for a unit of meaning, which is represented in speech by the intonational coloring of the constituent words.

The core structure of the synthesiser must therefore represent a sequence of words, and allow specification at various levels for their phrasing, syntactic features, semantic and pragmatic features, syllabic features, phonemic components, prosodic features, etc. This requires links within and between the levels for filling and following various relations, that allow for different degrees of abstraction or specificity in the representation of each.

Since it is fundamental to our philosophy of speech research that every component should be automatically replicable, we rely on large databases of sufficiently labelled speech to train the individual modules that fill in such information. These modules interact with language-dependent or speaker-

[1] *Chatr:* Collective Hacks from ATR

dependent information tables but are designed to be generic so that addition of new languages or speakers can be done without rewriting any synthesiser source code. The outer component of the Chatr core therefore consists of a means of interfacing to the external modules that read from (in synthesis) and write to (in training) the speech databases.

There is nothing in the core of Chatr that defines a language or a voice. Most human languages share the common features of higher-level abstract representations such as words and prosody, even if all do not use speech to realise them[2]. In many cases the mapping between feature-representations such as words, and their acoustic representations in speech is two-way and the core utterance structure can be as useful in creating labelled speech databases as in creating synthetic speech from more abstract representations of the utterance.

3 Speech corpora – basics

Traditional synthesisers tend to use very limited inventories of source units, relying on parametric modification to provide the prosodic variation. As such, they are constrained to producing only impoverished replications of natural human speech. By using larger corpora as sources of units for concatenative synthesis, we are able to reproduce more of the inherent variability of the original voice without resorting to artificial simulation. The richer the source corpus, the richer the speech synthesised using units taken from it. If the corpus is large enough and suitably labelled, then by incorporating prosodic information in the definition of the units to be selected from it, we can concatenate segments without the need for subsequent signal processing that tends to degrade the natural quality of the speech [8].

In selecting a corpus as a source for synthesis, we rarely know in advance what utterances we will be required to produce; if the synthesiser is simply to read texts or announcements out aloud, then the most appropriate source would consist of similar read material, but if it is to be part of a interactive conversational system, then read speech will be less appropriate. In order for the synthesis to sound spontaneous, the source must also have been spontaneous, in order for the prosody and voice quality to be in an appropriate range.

Early work using phonetically and prosodically rich 'balanced' corpora required reading sets of sentences that maximised variability and coverage of the language while minimising recording time and database size [12, 13, 14, 15]. We have found that this is not necessarily ideal. In producing these sentences, the psychological stress on the reader is frequently evident in the voice, as he or she struggles with the unusual combinations of sounds and the unnatural rhythms of the sentences constructed in this way. To create a natural-sounding friendly voice for conversational synthesis, we need the same relaxed qualities in our source corpus. We have designed the synthesiser to provide a set of tools for the processing of any arbitrary speech database as a source for synthesis. Because we cannot predict the ultimate usage environments, we label as many aspects of the speech corpus as we can, in the hope that the information acquired will be of subsequent use.

To create new utterances from the raw speech materials that are available in a given corpus, we select small, typically phone-sized, segments excised from prosodically and phonetically suitable contexts [5, 6]. By concatenating these, we can then recreate speech that not only has the same characteristics as

[2]In theory there is nothing to prevent output of sign-language or braille from the core representation, but in practice neither of these components has yet been added.

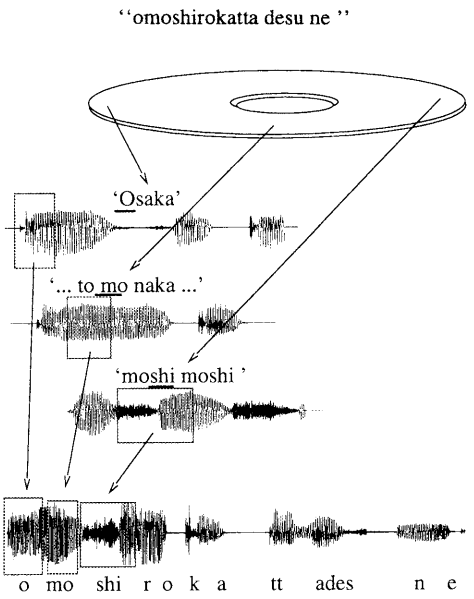

Fig. 1 Concatenation of segments for synthesis. Units are selected by prosodic as well as phonemic context.

the source speaker, but also keeps the characteristics of the original speaking style. In many cases, even if a given target sentence can be found verbatim in the corpus, the probability is high that its prosody would be inappropriate for the required context, and segments from other sentences will be used to reproduce it. Figure 1 illustrates this selection process.

To label such a corpus, we first produce an index into it, identifying each phone-sized speech segment both by its location in the parent waveform and by its phonetic features. Then to this basic index, we add the prosodic and voice-quality features by which the segments can be more precisely characterised. Since the labelling is largely automatic, we trade accuracy for volume; there will almost certainly be mis-labelled segments in an auto-labelled corpus, but we can rely on robustness in the subsequent selection procedures and redundancy in the labels to guard against this.

The quality of the resulting synthesised speech will depend on the variability and range of voice quality and prosody in the corpus. We have found from experience that the best synthesis comes from relatively steady continuous readings of a related texts, such as a short story. However, it is common in such recordings to make use of different voice qualities to portray different characters or moods, so at synthesis time, we have to differentiate amongst these to ensure that any given utterance will be produced with only the 'voice' of one character or mood. The labelling of 'affect' is still the subject of keen ongoing research.

4 Portraying meaning

The process of synthesis (and indeed speech in general) can be considered as two separate but interrelated information channels - text and voice. The former is primarily language-dependent, speaker-independent, and the latter to a large extent language-independent but speaker-dependent (the vocal-tract being common to all humans but unique for each). Their interaction is in the realm of meaning – the same text in a different voice will convey different interpersonal information, especially if coming from the same speaker. Because in principle any speaker could be required to produce any text, in any style, and in any language, these cross-dependencies require specific filtering.

4.1 Text-specific processing

A multi-lingual text may contain words of several languages, sometimes using different character encodings, for which the letter-to-sound mappings are language-dependent. However, we can not expect a Japanese

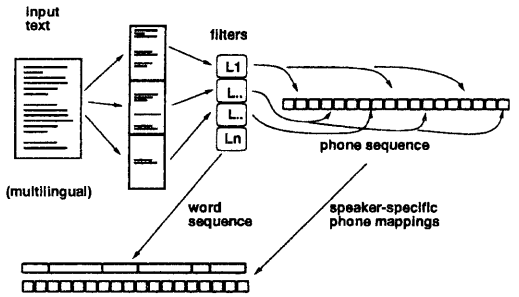

Fig. 2 Flow of linguistic information in the text processing stage. Currently for multi-lingual texts, four languages are supported.

letter-to-sound module, for example, to properly predict the pronunciation of either a German place name (possibly including umlauts, which coming in electronic form can appear as garbage unless special fonts are recognised) or a Korean street address possibly written in Hangul. Many text pre-processors will be confused by files which contain mixed character encodings, yet these are increasingly common in text files, especially in e-mail.

To overcome this problem we have designed language-specific filters for parsing and dictionary access so that each separate subsection of such mixed input can be independently processed and then, in a separate stage, we convert the time-aligned output of each filter to the nearest equivalent phone specification for the chosen speaker's phonetic inventory. A schematic for this shown in figure 2. Phone conversion requires a weighted bit-comparison of the phonetic feature specifications between the various source languages and those used in the index to the source corpus, a speaker-specific characteristic. These form the initial representation of the desired utterance in the synthesiser core.

In this way we can enable a Japanese speaker, for example, to pronounce Korean or English texts, but the pronunciation will be naturally limited to only those sounds available in the speaker's native language. For all mapping pairs we have tried so far (J-K-E-G) comprehensible renditions of the texts have been produced, but inevitably, the prosody was not that of a native speaker.

4.2 Voice-specific processing

Figure 3 illustrates some of the kinds of information represented in the core that determine the way an utterance can be produced. Some contributing factors such as speaker's mood, or environmental noise, are outside the scope of our synthesis, but we need to model much more than just the syntactic structure of the text to produce a natural-sounding and meaningful rendition of the speech.

To represent that difference in intended interpretation, we need rich information about the context and content of the utterance and a way of mapping between this and the desired acoustic characteristics of the speech. Although different from speaker to speaker, and from language to language, such acoustic features can be considered as implicitly coded in any given speech corpus, and by normalising them to an abstract specification level, we no longer have to predict them directly as absolute parameter values.

Depending on the nature of the input, we may have to fill in different parts of the core at different times. For example, if a target pitch contour and durations are already available, e.g., to mimic a given speaker's way of pronouncing an utterance, then the prosodic module need not be called, but if we have only the word string as input, then a cascade of processes will need to be invoked until we have enough information to predict a suitable intonation contour. If information about the focus relations is available, then

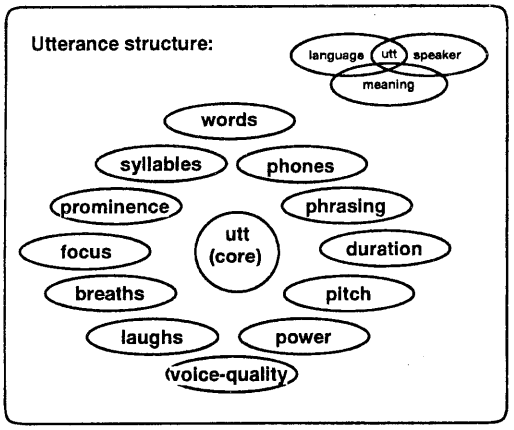

Fig. 3 Types of utterance-specific information in the synthesiser core. The symbol at the top right of the diagram reminds us that an utterance can be considered as the intersection of a subset of the language, as expressed by a given speaker to express a meaningful interpretation, which can be different for the same sequence of words.

extra processing can be performed to modify the prosody accordingly.

The core of the synthesiser forms a representation of each utterance and encodes various aspects of its content as a relational database. Some relations are hierarchical, as for words, syllables, and phones, but others are one-on-one (such as phones and their durations) or overlapping and time-aligned (words and laughter, or phrases and focus).

In order to select an appropriate candidate segment from those available in the speaker's corpus, we require a specification of the phone sequence using the labelling of the chosen speaker, and of the prosodic context for each phone, either as the associated durations, pitch, and amplitude targets or as an equivalent specification of prosodic context such as a ToBI labelling [17, 4]. From this information we are able to capture many of the finer voice-quality differences that characterise the speaker's speech and that we are as yet unable to specifically predict.

Breaths provide a good example of this finer selection using contextual information. Breathing is a fundamental human process, and speech cannot interfere with it much, but it should not be simply dismissed as noise as far as human vocal interaction is concerned. Utterance initially, we often hear a sharp intake of breath, which can cue the listener that a speech signal is coming. However, in our synthesis, it is not necessary to specify such breaths explicitly, they appear naturally as part of the pause that is taken from an appropriate inter-sentential context, and will be selected by default when the specification of database units is in prosodic as well as segmental terms. The same is true of sentence-final pitch-fall, power-drop, and glottalisation. When the speech segments are selected from an appropriate prosodic context, they will have the required phonatory characteristics naturally.

5 Choices

As a result of the modular design, the user of the synthesiser is presented with a choice of options or 'switches' that define the routing of processing for any given utterance. There are default specifications for every stage, but at the most basic level of interaction one can simply specify a text file in order to hear it spoken. More interesting uses require selection of various features, to determine the route taken in generating an utterance from the information in the core.

First, to load the core structure, input can be made manually from the Unix command-line, read in from a file or an emacs buffer, or accepted via sockets or a Java interface from the Internet. Input need not necessarily be text, it might be by voice, to modify or cor-

rect an inappropriate prosodic specification, or via a graphic interface to modify a pitch contour, for example. It might be at a high abstract level, as words alone, or more fully defined as specific parameter values for the individual components.

The most fundamental 'switch' concerns the choice of voice or speaker. By selecting a different speaker, not only the voice quality, but also the language or speaking-style will undergo changes, as the characteristics of a different database will be used to produce the utterance. On selecting a speaker, many of the defaults will change automatically. For example, a database of Japanese will typically have been labelled using a different phone set than those used for Korean or English. Within a language, dialectal differences may require use of a different dictionary. Such settings are specified as 'speaker defaults'.

In addition to these implicit parameter switches, there are also 'methodological switches' that allow the replacement of one prediction module with another, for example, to opt for the 'Fujisaki' method of intonation prediction rather than the 'ToBI' one, or to compare syllable-based duration predictions with phone-level predictions, and to specify different waveform-generation methods or audio output devices. Being a research 'toolkit', there are new options always being added.

6 Future directions

Several aspects of meaning and interpersonal cues have yet to be fully expressed in Chatr. In previous work we have shown that prosodic cues have clear correlates in speech-act differences for the same set of words [1]. If we know for example that the Japanese phrase 'sou desu ne' is intended as a hesitation, then we can synthesise it with very different segments than would be used if it were intended to signal a confirmation. In a speech translation environment, such information is available with the word-string to be synthesised.

6.1 Extra-linguistic noises

All human speech includes disfluencies, and speakers often make errors. A common reaction by the speaker to such an event is to laugh. Since our unit-selection is by weighted cost criteria, we have a measure of how good any particular set of segments is in representing the speech it is supposed to synthesise. It is only a small step to make the synthesiser laugh apologetically when it makes a poor performance. However, the ways of using such meta-linguistic signals still need much research.

Emoticons (now commonly used in e-mail :-) have evolved to express non-linguistic information such as smiling, sadness, and laughter, that play an essential part of interpersonal communication. Laughter can signal to a listener that no offense is intended in cases where the wording may be ambiguous; in informal conversations, sniffing can be used to express sadness, and in Japanese a hissing intake of breath through the teeth signals disapproval. Synthesis must also learn to produce such noises if it is to be more widely used.

The sounds commonly exist and can be labelled just as other phonemic sounds can. In one database of half an hour of speech we counted 163 laughs, and were able to categorise them into 7 varieties. Now we have a laughing synthesiser.

6.2 Finally, Is stereo better?

In theory, since a human speaker has only one mouth, a monaural sound source should be adequate for speech synthesis, but in practice, we have had very interesting results from synthesis using waveforms from stereophonic

recordings. In terms of implementation in the synthesiser there is negligible difference between a mono and a stereo source, since the index into the speech is by time alone, and it is only the audio equipment that performs any demodulation. Sampling rate and waveform complexity are independent of the synthesis process.

That stereo speech synthesis should sound better than mono offers evidence for the attention to such details as differential airflow through the nostrils, or differential radiation from the mouth according to laterality of the tongue. Since most domestic audio equipment is now stereophonic, perhaps synthesis should conform here too. Chatr can do it, naturally.

7 Conclusion

This paper has described a new approach to speech synthesis, that offers an engine for processing existing speech databases rather than a machine that just encodes a voice. Under this method, we conserve the characteristics of speaker and style, and open up a new level of naturalness in synthetic speech. As larger corpora become available, we can expect more spontaneous speech phenomena to be modelled, and can thereby envisage uses for synthesis that were not feasible when it was restricted to just reading texts.

References

[1] 藤尾 茂・ニック キャンベル・樋口宜男："韻律を用いたテキスト非限定型発話アクト識別方式", 日本音響学会春季研究発表講演論文集, pp. 245-246 (1996).

[2] G. Bailly and C. Benoit, eds, *Talking Machines*, North-Holland, 1992.

[3] T. P. Barnwell, "An algorithm for segmental durations in a reading machine context". Technical Report 479, MIT, - Research Laboratory of Electronics. 1971.

[4] M. E. Beckman & G. M. Ayers. Guidelines to ToBI Labelling. Version 2.0., Ohio-State University Technical Report, 1994.

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

[6] 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.

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

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

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

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

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

[12] N. Iwahashi, N. Kaiki, & Y. Sagisaka, "Concatenative speepagesch synthesis by minimum distortion criteria", pp65-68 *ICASSP '92*.

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

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

[15] 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.

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

[17] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling English prosody, 867–870, *ICSLP92*.