Procesamiento del Lenguaje Natural, num. 29 (2002), pp.165-172 recibido 28-05-2002; aceptado 04-07-2002

# Towards More Natural Synthetic Speech

**Pilar Manchón Portillo**
Universidad de Sevilla
pmanchon@yahoo.com

**Abstract:** This article reports the results of two experiments in which factors such as duration, amplitude and noise are manipulated, in order to achieve more natural utterances in synthetic speech. The participants were native speakers of English, instructed to judge the naturalness of the different versions of utterances generated throughout the manipulations. The results indicate that there are significant individual preferences, as well as classification principles other than conventional ones. There is evidence to believe that further research in this area will render positive results in the search for naturalness. The same principles could be applied to search for naturalness in the prosodic structure of the synthetic utterances. Advancement in this area will surely render improvements in Spoken Dialogue Systems.
**Keywords:** Naturalness, Synthetic Speech, Human-Computer Interaction, White Noise, Duration, Entropy, Spoken Dialogue Systems.

## *1 Introduction*

Natural sounding synthetic speech is one of the goals in language research.

Although it is one of the main challenges in Language Engineering, 'Naturalness in synthetic speech is a very subjective aspect that cannot be easily measured. Native speakers can easily tell whether something sounds natural or not; what is not always so easy to define is why.

In the experiments reported here, different versions of synthetically generated words were presented to a number of native speakers of English. They were instructed to rank these versions in terms of naturalness. Each set of versions contained the default generated by the synthesizer (from the Boston Radio Corpus), and a number of hand-manipulated versions.

Firstly, we must decide what manipulations and in what degree they should be performed. Secondly, we must choose the principle or rule to perform these manipulations systematically.

The manipulations chosen were segment duration reduction, introduction of white noise and amplitude reduction. The use of these factors is justified by the results obtained in other studies (Hoequist 1983; cited by K. Tajima et al. 1997). The points at which these manipulations have been applied as well as the manner in which it was done were selected considering two main concepts: the recognition point, and the maximum and minimum entropy points.

### 1.1 Interpreting Speech

Since speech is continuous, the process of interpreting speech is also continuous. Moreover, the phonemes that make up utterances overlap, making the segmentation process much more difficult. Research in this area confirms that there is a continuous mapping of speech onto the lexical, semantic and pragmatic levels of interpretation, at the maximum speed possible. However, this does not mean that the interpretation of messages occurs as the speech signal is perceived, at least not in all cases.

Experiments in this area (Bard, Shillcock & Altmann, 1987) prove that 20% of words are not recognized by the time the whole word is heard and it is more likely to affect function than content words.

One of the relevant concepts is the notion of redundancy. It is the existence of a great deal of redundancy in speech that allows for the reductions in duration and amplitude that seem to occur in natural speech.

Part of the controversy in the unravelling of the selection process is the interdependence between form and content; which of them is prioritised in the selection process is still undetermined. Reaction-time experiments indicate that listeners can recognize words before hearing their ending, if all possible candidates have been discarded at an earlier stage. Marslen-Wilson's concept of 'early selection' (Marslen-Wilson 1987) may explain

ISSN 1135-5948 

the speed of speech interpretation and shed light on the interrelation between form and concept:

> *Early selection is the identification of spoken words, in normal utterance contexts, before sufficient acoustic-phonemic information has accumulated to allow the identification decision to be made on this basis alone. Numerous studies (...) show not only that words are, on average, recognized on context about 200 msecs. from word onset, but also that the sensory information available at that point is normally quite insufficient by itself to allow the correct identification of the word being heard*.

The system must be able to combine the processes of access and integration that define the mental lexicon as an information processing system. Multiple access and multiple assessment of integration must then be simultaneously possible. The speed with which this process takes place indicates some form of parallelism in the performance of the different tasks.

This information is extremely important in the design of Spoken Dialogue Systems, where the ultimate goal is to simulate the human brain speech processing abilities.

## 1.2 Lexical Process

There are a considerable number of approaches that attempt to describe the lexical process. In general three basic assumptions have been agreed on in terms of lexical access and selection:

**a**) The concept of activation is an accurate representation of the process.

**b**) A set of candidates is simultaneously activated and competes in the selection process.

**c**) The levels of activation determine the selection of the final candidates.

There are also some controversial issues that have not yet been agreed upon (Bard, 1990):

**a**) The frequency of competitors affects the activation level of a word.

**b**) Competition among candidates involves lateral inhibition.

**c**) The perceptual choice criterion is determined by the ratio of activation of a candidate to the total activation, or to its closest competitors.

The recognition process could then depend on the level of activation of the words and their competitors. Furthermore, assuming that frequency affects activation levels in lexical access, we may assume that fluctuations in frequency also affect the interrelation between each individual item and its competitors in terms of the activation levels (Marslen-Wilson 1990).

Although the other models of lexical access will not be discussed, we must point out that it is Marlen-Wilson s Cohort Model that has been essential in the design of these experiments, since it allows for the prediction of the point at which a certain word will be recognized. This model also implies continuous and sequential lexical access and selection over time.

An alternative perspective of how the information reaches the lexicon is given in terms of Entropy (H), i.e., the amount of information conveyed by the subsequent segments of a given word.

It will be on these two concepts, *recognition* or *uniqueness point* and *Entropy* that we will base the manipulations performed for the experiments described below.

The Cohort Model presents as well some problems. Bearing in mind that the word-initial cohort is defined in terms of the beginnings of the words, the essential information on which the model bases the entire lexical decision process, must be the word onset.

Once the word-initial cohort is defined, no other candidates can be considered or included in the decision set. Therefore, this model cannot cover those cases defined by Bard, Shillcock & Altmann 1987. The nature of speech itself makes it unlikely to guarantee the correct estimate of the word onset, in which case, the recognition process is doomed to failure.

## 1.3 Semantic Access

There is significant evidence for the multiple activation of lexical contents and phonological elements early in the access and selection process (Marslen-Wilson et al., manuscript; Zwitserlood 1985; Lucas 1987; Seidenberg, Tanenhaus, Leiman & Bienkowski 1982). Given that this implies the use of sensory and contextual constraints interrelated, the interaction of these constraints in terms of timing and manner is significantly relevant.

Some models (Logogen Model, Morton 1969) propose the existence of *contextual preselection*. Forster supports a double-mode system where the lexical access module can be

tuned in form-based bottom-up mode by default. The Cohort Model however, advocates for a bottom-up priority process which presupposes no context-based preselection, and appoints the form-based selection process as the determining factor. There is extensive evidence confirming that strong contextual constraints do not prevent the activation of semantically inappropriate candidates that nonetheless match the auditory input ( Tyler 1984; Tyler & Wessels 1983; Samuel 1981; Zwitserlood 1985; all cited in Marslen-Wilson 1987). There are therefore strong arguments to believe that contextual constraints contribute to the selection process only when the form-based selection process has already appointed one candidate as the most likely.

## *2 Experiment 1*

### 2.1 Subjects

Twenty naive subjects, all of them native speakers of English, completed the experiment.

### 2.2 Synthesizer

The synthesizer used was the Festival Speech Synthesis System: version 1.4.0. The voice chosen was Kurt, an American English male speaker. It uses the UniSyn residual excited LPC diphone synthesizer. This uses the CMU lexicon, and letter to sound rules trained from it. Intonation is trained from the Boston University FM Radio corpus. Duration for this voice also comes from that database.

### 2.3 Items

Sixty words were randomly selected from the CELEX Lexical Database: 5 from each of 8 word class groups and 20 more solely because of their unusual length. Ten of these words were longer than 16 characters, and the rest at least 12 characters long.

### 2.4 Objective

The goal of the experiment was to present the subjects with a number of manipulated versions of each of the words, which were synthesized in isolation. The subjects were instructed to rank them in terms of their ‘naturalness .

### 2.5 Manipulations

Depending on their length, data availability and other factors, the items selected for the experiments were classified in several groups. For each group only a number of transformations were applied in order to generate a limited number of versions. The following groups have been analysed separately:

| Group | Description |
|---|---|
| L | Word Length >= 16 |
| 8V | 8 Comparable Vs. |
| 6V | 6 Comparable Vs. |
| 5V | 5 Comparable Vs. |
| 2V | 2 Comparable Vs. |

The following table shows all the manipulations performed in general.

| Labels | Description |
|---|---|
| D | Default |
| 20 | Duration 20% |
| 40 | Duration 40% |
| P20 | Duration Progressively 20% |
| P40 | Duration Progressively 40% |
| A | Amplitude 0.2 |
| PA | Amplitude Progressively 0.2 |
| Plus | Noise Added |
| H20 | Entropy 20% |
| H40 | Entropy 40% |
| HA | Entropy Amplitude 0.2 |

#### 2.5.1 Recognition Point

##### *2.5.1.1 Duration*

The duration of the individual phonemes within the word was reduced from the recognition point on.

In the case labelled ‘20 , all phonemes after the recognition point were reduced in length 20% of their default duration.

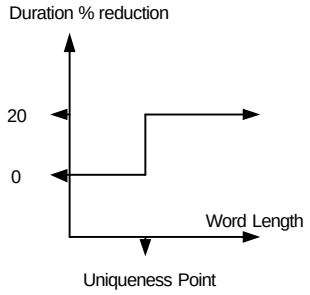


Fig1: Duration reduction

In ‘40 , the manipulation is identical to ‘20 except for the percentage of reduction, which in this case is 40%.

In ‘P20 , the phonemes are reduced progressively up to 20%, and up to 40% in ‘P40 .

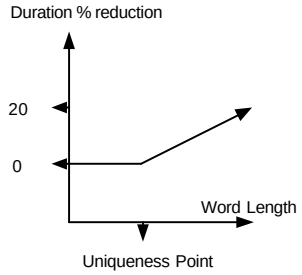


Fig2:Progressive duration reduction

*2.5.1.2 Noise*

The introduction of white noise is inspired on results obtained in vision studies, where a certain amount of blur was found natural in the perception of natural images. This will give us an idea of the amount of noise bearable for the listeners, or even preferred in terms of naturalness. In ‘Plus , white noise at a scale of 100 was created an added to the word from the uniqueness point on.

*2.5.1.3 Amplitude*

In ‘A and ‘PA the amplitude of the phonemes has been reduced in a step function manner in ‘A and progressively in ‘PA .

### 2.5.2 Entropy

This is virtually a measure of the amount of information provided by each phoneme in relation with the total number of phonemes in the word.

The versions generated in terms of entropy are based on the data obtained in the MOP Project at the University of Edinburgh

*2.5.2.1 Duration*

In ‘H20 and ‘H40 the phoneme with the minimum entropy value once the entropy peak has been reached was reduced 20 and 40 percent respectively, and the rest of the phonemes from the peak on were reduced in a percentage inversely proportional to their entropy value (MOP project, Edinburgh University).

*2.5.2.2 Amplitude*

In ‘HA , the amplitude of the phonemes is reduced in terms of their entropy.

## 2.6 Manipulations

Due to the nature of the data and given the difficulty of consistently ranking sets of seven and eight items, the analysis of the results will be based on the number of times that each of the versions has been ranked first or second, rather than on the mean scores. The results of both approaches will be compared.

The versions will systematically be ordered in terms of the highest means, and the most significant relations will be drawn and labelled according to the following scheme:

- $p < 0.01$ labelled as “>>”.
- $p < 0.05$ labelled as “>”.
- $p >= 0.05$ labelled as “=”.

Fig. 3 displays the differences in L according to Data Analysis 1 (DA1), where versions have been ordered according to the number of times they have been ranked first.

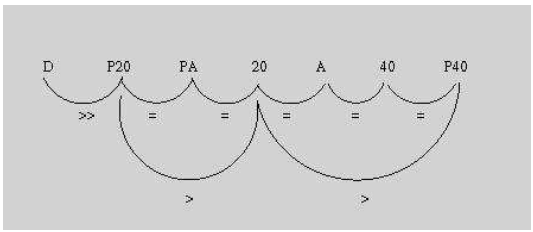


Fig 3: Group L, DA1

Fig. 4 shows the same results but according to Data Analysis 2 (DA2), where the versions have been ranked according to the mean number of times they have been ranked first or second.

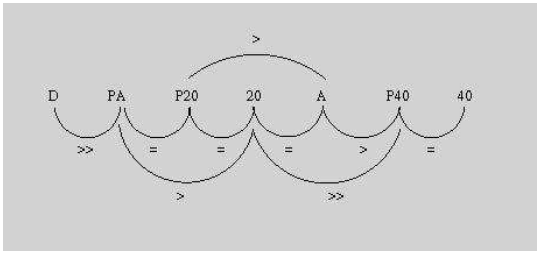


Fig 4: Group L, DA2

There are significant differences in the order of preference and the statistical level of significance. Both approaches will be used for the groups where the number of versions is high. All other groups were analysed in the same fashion. The results for groups 6V and 2V were not statistically significant for this analysis.

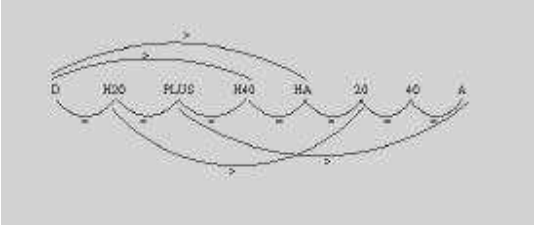


Fig 5: Group 8V, DA1

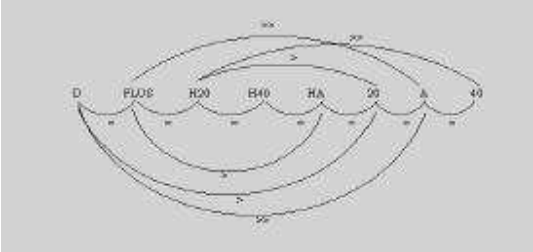


Fig 6: Group 8V, DA2

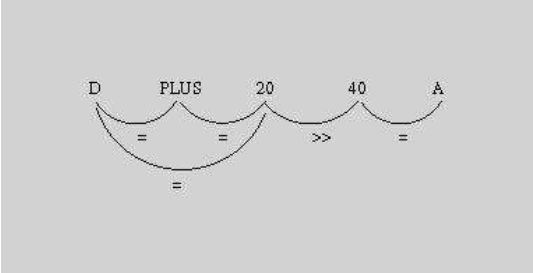


Fig 7: Group 5V, DA1 and DA2

The analyses presented show the relationship between the versions, given the choices of all subjects. However, there could exist patterns within subjects. With this goal in mind, part of the data was analysed and the following results were found:

| **Groups** | **Analysis of Variance (ANOVA)** |
|---|---|
| L (DA1) | p = 0.000 |
| L (DA2) | p = 0.000 |
| 8V (DA1) | p = 0.000 |
| 8V (DA2) | p = 0.000 |
| 2V (DA1) | p = 0.068 |

The statistical significance above indicates that the subjects consistently chose the same versions throughout the experiment.

## *3 Experiment 2*

### 3.1 General Conditions

The resources used for this experiment are identical to Experiment 1.

The selection of versions for Experiment 2 was based on the preliminary analysis of the results of Experiment 1. Consequently, the new group can be classified in three sub-sets.

| **Vers** | | | | | **Items** |
|---|---|---|---|---|---|
| L | D | 20 | P20 | PA | 20 |
| F8V | D | 20 | H20 | HA | 10 |
| F5V | D | 20 | 40 | A | 5 |

### 3.2 Objective

The goal of this experiment was to modify some of the factors that were likely to have some effect in Experiment 1, such as the high number of versions presented. To ensure short term memory, in Experiment 2 we have reduced the number of versions of each group to 4, and the number of words to 35. Only DA1 will be performed.

### 3.3 Results

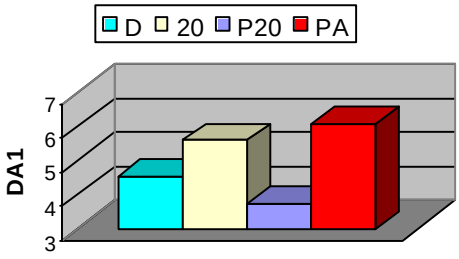


Fig 8: L in Ex2

The maximum difference in L is that between PA and P20, and the level of significance is ***p < 0.07.***

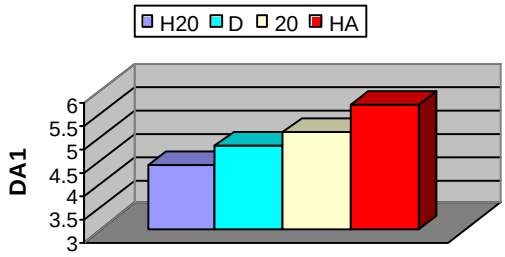


Fig 9: F8V in Exp. 2

The maximum difference in 5V is that between HA and H20. Its level of significance is $p < 0.7$.

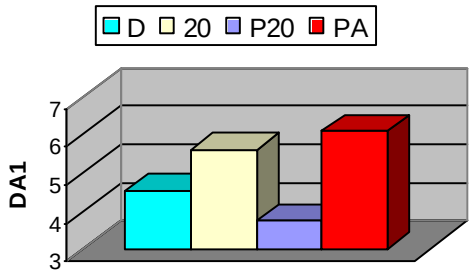


Fig 10: F5V in Exp. 2

In F5V, at least two of these versions are significantly different: $p < 0.034$.

## 4 Conclusions and Future Work

Given the small number of subjects, statistical significance is unlikely here. Nonetheless, the goal of these experiments is to find tendencies and relationships, rather than irrefutable proof of facts, which would require more extensive experiments.

In most cases the extreme manipulations were deemed to be worse than the rest and therefore discarded. However, there is a great number of possibilities regarding the combination of manipulations. It would be interesting to generate a version manipulated both in terms of duration and amplitude.

Since the base-form used was generated from observational data, it was unlikely that simple manipulations would render better results overall. However, a relationship between word frequency and the quantity and quality of manipulation seems plausible. This opens more possibilities for further research.

With regard to the subject analysis, 7 out of 10 subjects present significant version preference. No relationship between the individual subjects and their preference has been so far established. Their preferences do not appear to be related to their variety of English. Further analysis is required.

| Version Preference L | | | | |
|---|---|---|---|---|
| **Exp 1** | D | P20 | PA | 20 |
| **Exp 2** | PA | 20 | D | P20 |
| **Version Preference 8V** | | | | |
| **Exp 1** | D | H20 | HA | 20 |
| **Exp 2** | HA | 20 | D | H20 |

Fig11:Preference in L and 8V in Exps. 1 and 2

### 4.1 Conclusions

These results point towards the kind of manipulations that could help improve naturalness in synthetic speech. Although the results are not conclusive, they establish a basis for further experimentation.

The number of versions presented seems to have an effect on the decision capacity of the participants and, although these conclusions need confirmation, ignoring such factor might invalidate future results.

New principles of classification are likely to offer new perspectives in this area. Therefore, features like frequency, word class, number of phonemes, entropy, number of phonemes after the uniqueness point and stress are very likely to provide further information with regard to the possible manipulations. Although these interrelations are likely to be more complex, the analysis of simple combinations is necessary to establish more complex connections.

One of the most striking findings is the possibility of the existence of personal preferences regarding longer words. Moreover, these preferences do not appear to be related to English variety, age or sex. It is then possible that this difference in preferences could conform to a certain pattern, or be related to a certain factor, even though neither of these have yet been identified. Most systems are designed to be used by a great variety of individuals but there are cases however, in which one

individual uses the system for long, continuous periods of time, or very often. It should then be possible to adapt the system. This adaptation should be easy to program, or even automatic in Spoken Dialogue Systems.

### 4.2 Future Work

It would be of great interest to implement the results of this research in spoken dialogue systems, where additional information is available. As we have seen in the introduction, much information in addition to the acoustic information alone is used in the understanding process. Spoken dialogue systems could make use of all this information.

According to the approach chosen here, all these words have the same recognition point. Further research in this area is certainly needed. It is by no means obvious that a native speaker presented with one of these different cases:

| / ‘dQgm .../ | / dQg ‘m.../ |
|---|---|
| *as in* | *as in* |
| / ‘dQgm@ / | /dQg ‘m{tlk / |
| / ‘dQgm@tIz@m / | / dQg ‘m{tlkP / |
| / ‘dQgm@tIst / | /dQg ‘m{tlks / |

Within the same paradigm of words:

| |
|---|
| dogmas |
| dogmatic |
| dogmatical |
| dogmatically |
| dogmatics |
| dogmatism |
| dogmatist |
| dogmatists |
| dogmatize |
| dogmatized |

Would consider them both as a viable possibility. Whenever one of them is heard, the whole paradigm could be primed, but whether that means that some of the members of the paradigm are considered as competitors even though their stress pattern is different, is not straightforward.

The most recent approaches to lexical access consider frequency a key factor in lexical competition. Researchers have not yet agreed on what the right approach is in terms of the frequency of the words themselves, and the frequency of their competitors. Therefore, more experiments relating these factors will contribute to solving the puzzle.

## *5 Bibliography*

Altmann, G. *Lexical Statistics and Cognitive Models of Speech Processing* in Altmann, G. (Ed) Cognitive Models of Speech Processing. MIT Press Cambridge: MA, 1990.

Bagshaw, P. *Phonemic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression*. Computer Speech and Language (1998) 12, 119-142.

Bard, E. *Competition, Lateral Inhibition, and Frequency: Comments of Chapters of Frauenfelder & Peeters, Marslen-Wilson, and Others* in Altmann, G. (Ed) Cognitive Models of Speech Processing. MIT Press Cambridge: MA, 1990.

Bard, E & Shillcock, R. *Competitor Effects During Lexical Access: Chasing Zipf s Tail* in Shillcock, R. & G. Altmann (Eds). Cognitive Models of Speech Processing: The Second Sperlonga Meeting. Lawrence Erlbaum Associates Ltd. UK: Hove, 1993.

Butterworth, B. *Lexical Access in Speech Production* in Marslen-Wilson, W. (Ed.) Lexical Representation and Process. MIT Press. Cambridge: MA, 1989.

Charles-Luce, J., Luce, P. & Cluff, M. *Retroactive Influence of Syllable Neighbourhoods* in Altmann, G. (Ed) Cognitive Models of Speech Processing. MIT Press Cambridge: MA, 1990.

Cutler, A. *Auditory Lexical Access: Where Do We Start?* in Marslen-Wilson, W. (Ed.) Lexical Representation and Process. MIT Press. Cambridge: MA, 1989.

Damper, R., Marchand, Y., Adamson, M. & Gustafson. *Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches.* Computer Speech and Language (1999) 13, 155-176.

Fourcin, A. *Assessment of synthetic speech* in Bailly, G., Benoît, C. & Sawallis, T. (Eds). Talking Machines: Theories, Models and Designs. Elsevier Science Publisher. Amsterdam: The Netherlands, 1992.

Griffin, Z. & Bock, C. *Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production*. Journal of Memory and Language 38, 313-338 (1998).

Kessinger, R. & Blumstein, S. *Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies*. Journal of Phonetics (1998) 26, 117-128.

Klatt, D. *Review of Selected Models of Speech Perception* in Marslen-Wilson, W. (Ed.) Lexical Representation and Process. MIT Press. Cambridge: MA, 1989.

Manchón Portillo, P. *Psychokinetically Inhibited Manoeuvrability: Towards More Natural Synthetic Speech*. MSc. Thesis at Edinburgh University, 1999.

Marslen-Wilson, W. *Access and Integration: Projecting Sound onto Meaning* in Marslen-Wilson, W. (Ed.) Lexical Representation and Process. MIT Press. Cambridge: MA, 1989.

Marslen-Wilson, W. *Issues of Process and Representation in Lexical Access* in Shillcock, R. & G. Altmann (Eds). Cognitive Models of Speech Processing: The Second Sperlonga Meeting. Lawrence Erlbaum Associates Ltd. UK: Hove, 1993.

Pecher, D., Zeelenberg, R. & Raaijmakers, J. *Does Pizza Prime Coin? Perceptual Priming in Lexical Decision and Pronunciation*. Journal of Memory and Language 38, 401-418 (1998).

Shillcock, R. & Bard, E. *Modularity and the Processing of Closed-class Words* in Shillcock, R. & G. Altmann (Eds). Cognitive Models of Speech Processing: The Second Sperlonga Meeting. Lawrence Erlbaum Associates Ltd. UK: Hove, 1993.

Segui, J. & Grainger, J. *An Overview of Neighbourhood Effects in Word Recognition* in Shillcock, R. & G. Altmann (Eds). Cognitive Models of Speech Processing: The Second Sperlonga Meeting. Lawrence Erlbaum Associates Ltd. UK: Hove, 1993.