Proceedings of the Institute of Acoustics

TOWARDS HIGH QUALITY AUTOMATIC ANNOUNCEMENT SYSTEMS

T. J. Gillott, M. C. Hall, S. Macgregor
British Telecom Research Laboratories,
Martlesham Heath,
Ipswich IP5 7RE.

ABSTRACT

Concatenated speech offers potentially very high quality output for speech synthesis systems. Such quality is very desirable for automatic announcement systems which need to put out variable content messages over the telephone network. However, to date the full potential of the techniques has not been fulfilled, largely as a result of the difficulties experienced in producing a natural pitch contour, and other prosodic problems.

This paper outlines these principal problems, and describes various approaches which are being investigated to alleviate them. Possible solutions are applied to a small well-defined domain, which is the automatic generation of telephone numbers. A phonological intonation model is developed using formant-encoded speech templates. Features of existing and projected systems are discussed.

1 GENERAL

Any automatic message generation system working over the Public Switched Telephone Network must produce speech of sufficient quality to be fully acceptable to the general public. To date, BT has used several versions of telephone number announcement systems. While such systems produced successively higher speech quality, poor prosodics limited overall naturalness.

This paper addresses the problems of prosodics in synthetic spoken telephone numbers. The two main areas examined are F0 and timing. A systematic study of the prosodic behaviour of digits in spoken telephone numbers was undertaken. Algorithms were derived from the results of this study and implemented in software.

This paper presents an overview of the architecture of the resulting software, termed ANSA (standing for Automatic Number Synthesis and Announcement), which was used as a testbed for various prosodic techniques. ANSA uses a combination of a phonological intonation model and timing calculation techniques to synthesize high-quality spoken telephone numbers.

The problems facing automatic announcement systems are outlined below, along with some past and present techniques adopted by BT to cope with the problem. The latter are presented in the context of the ANSA software.

2 PROBLEMS

2.1 SPEECH QUALITY

Any automatic message generator should produce highly understandable speech which also has a reasonable degree of naturalness. An automatic telephone number generator can use various units, ranging typically from entire telephone numbers to single concatenated digits. In general, the larger the unit employed, the more natural the resulting number. Unfortunately, the larger the unit becomes, the more data has to be collected. Earlier systems attempted to reduce the data inventory by using the concatenated single-digit approach.

However, the penalty for using this limited digit inventory is that of reduced naturalness, particularly in the prosodic area. This has not prevented the technique from being used; the early announcement systems mentioned above used concatenated digits with fixed intonation patterns. This caused many problems for the user, despite the basic digits themselves being easily intelligible. This is discussed further in the section below on Intonation.

2.2 INTONATION

Intonation is the single largest problem faced when designing a system of this type. Although systems can use fixed intonation patterns with some success, the resulting concatenated digit strings are only marginally natural.

This type of system carries a disadvantage in that the lack of any temporal markers (most of which are carried by intonation) causes an unusually high demand to be placed on short-term memory function. As a result, the user finds it hard to process such numbers.

Most of the efforts to improve on automatic number announcement systems have focussed on the problem of intonation, to the exclusion of all else. A typical approach was to sample three forms of each digit, one from the end of the subscriber code (Terminating), one from the middle of a number group (Neutral) and one from the end of a non-final group (Continuant) [1].

System X announcements used PCM-encoded digits, and used the above method to simulate F0 patterns. The restrictions imposed by the waveform-resynthesis method necessarily limited the complexity of the model, but the resulting quality was a great improvement on previous systems.

The next stage was to use formant-encoded speech templates in place of PCM-encoded speech [2,3]. This gave a much greater flexibility in that, under this system, F0 is a single easily-accessible parameter that can be calculated and re-assigned. A system was produced [3] which re-implemented the System X intonation model, but which included two refinements.

The first was an additional template; it was found necessary to produce a "phrase initial" pattern. The second concerned adjustments of the timing of each individual digit to produce perceptual regularity. This last used an implementation of Morton et al.'s P-Centre notion [4]. This system produced reasonable quality speech with a fairly natural F0 contour, although it did not attempt to model any second-order phenomena such as vowel and consonant perturbations.

3 THE ANSA MODEL - SYNTAX, PHONOLOGY AND PROSODY

3.1 SYNTAX AND PHONOLOGY

3.1.1 SYNTAX

The first stage of ANSA is a parsing module which splits up the incoming digit string. To produce a sensible result, the syntax of the telephone number must be taken into account. A study of the hierarchical nature of the telephone number produced a simple syntactical model which could be used as a basis for an intonational model.

In this model, the number is considered as a set of syntactic Groups. There are two different types of Group. The distinction is based on syntactic and intonational behaviour. Following conventional syntactic terminology, these two Group types are referred to as the Head and the Modifier. The Head is the most important part of the number, both syntactically and semantically. It corresponds to the Subscriber Code. All preceding codes are regarded as subordinate to the Head; they modify it in the same way as adjectives can be regarded as modifying nouns. Modifier Groups correspond to Area Codes. Thus, the number 081-688-5673 will be split up by the parser as

081 688 5673
MODIFIER MODIFIER HEAD

The parser requires separators to inform it of the existence of more than one Group. Following traditional notation, these separators may be either dashes or spaces.

3.1.2 THE PHONOLOGICAL INTERFACE

The framework used for the generation of a pitch contour is an implementation of a theoretical Pitch Event model of intonation proposed in [5]. This framework contains two components, a phonological and a phonetic one, completely independent of each other. Neither the phonetic nor the phonological component of the framework make any assumptions about any other formant parameters. Thus the intonation module as a whole is completely independent of the rest of the processing. ANSA preserves and exploits this independence, allowing completely different types of formant encoded data to be used with minimal changes.

Forming the phonological interface are various rules which define the Pitch Events which are going to be assigned to the templates. A sub-module of the Concatenation Module is responsible for assigning these.

This sub-module defines all syntactic Groups as Major Phrases, with the important exception of the Head Group, which is complex and may consist of more than one Major Phrase. A Major Phrase is defined here as being an utterance terminated by a breath or punctuation.

Each Major Phrase is assigned a pattern of perceptual prominences, which are eventually interpreted as Hertz values. The rules governing these prominence assignments reflect the patterns of intonational behaviour exhibited by different numbers of stressed syllables in a phrase.

The phrase based prominence pattern assignments act as a direct input to the intonation model. Each Major Phrase is processed separately. Features of the intonation model are described below.

3.2 INTONATION GENERATION

The Intonation Modules are responsible for producing a train of F0 values from the phonological instructions passed to them by the Concatenation Module. Briefly, the basic contour is produced from frequency values in Hertz which are assigned to each Pitch Event. These are derived from the perceptual prominence values which are assigned in the phonological component of the Concatenation process. Interpolation is then performed between Pitch Events to produce a continuous contour. However, two additional stages of processing are required to produce a fully natural F0 contour; intrinsic F0 and consonantal perturbations.

3.2.1 VOWEL INTRINSIC FUNDAMENTAL FREQUENCY

For some time, vowels have been observed to contribute to F0 quality. Depending on the vowel, a smooth local perturbation will be produced on the F0 contour, resulting in either higher or lower F0 values in the region of the vowel itself. This is referred to in various studies [6,7,8 etc.]. as (vowel-) intrinsic fundamental frequency (IF0).

The present model uses an algorithm which treats IF0 as an integral part of the phonetic realisation process performed on the phonological Pitch Events. Each vowel has an IF0 contribution associated with it, which is combined non-linearly with the Hertz values assigned when a Pitch Event is realised. This produces the necessary smooth perturbation during the stressed vowel.

3.2.2 CONSONANTAL PERTURBATION

Consonantal perturbations are also observed in F0 traces. These perturbations are unlike those introduced by vowels in that they introduce sharp discontinuities into the pitch contour. ANSA contains a module which calculates appropriate perturbations for each consonant, and assigns these to the vowel at the appropriate positions.

3.2.3 INTONATION ANNOTATION

ANSA depends for its operation on a secondary data file which stores all the data for relevant events in a template. Each template entered into the main lookup table must have a corresponding entry in the annotation table which contains data such as duration, start of stressed vowel, etc. Using these data, it is possible to produce a complete description of the relevant intonation events throughout any word, and to tune prosodic information without recompiling the system.

4 THE P-CENTRE AND OTHER PROSODIC CONSIDERATIONS

In order to produce natural-sounding telephone numbers, it is not enough merely to produce a convincing F0 contour. The Groups mentioned above, when spoken by operators, are separated and spaced in a rather artificial way to enable the listener to process the number adequately. The prosodics of these separations and spaces are similar to that used when dictating words. An algorithm must be identified which models the rhythm and spacing of numbers correctly.

Initially, experiments were carried out with unperturbed F0. An algorithm for spacing groups was fairly readily identified. A first approximation was to use fixed silence periods of half a second. This was unsatisfactory, but a second approximation using the length of the Group itself as a metric for the following silence produced better results.

A further problem was that digits within the Groups appeared to be spoken at an irregular rate, some following others far too fast, others too slowly. This perception of unevenness was not apparently linked to the duration of the digit.

The problem of perceptual unevenness when concatenating words has been widely recognised. Many attempts have been made to space digits and other sounds such that they appear to occur evenly in time. Morton et al. [4] suggest that there is a "perceptual centre" to every word, which must occur at regular intervals. Producing an algorithm to implement this notion was initially attractive given the ability to annotate any part of the digit template, but on investigation a significant problem arose. This was that the Perceptual Centre, or "P-Centre", did not correspond with any one acoustic event within the syllable. Various intra-syllable events were tested, but none of these produced consistent results.

Subsequent tests on perturbed F0, ie F0 with vowel and consonant perturbations added to the basic contour, produced an interesting result. Much of the previous unevenness disappeared, and the number strings were perceived as having much more natural timing, to the point that it was felt that no more experimentation on P-Centres was necessary.

This has implications for perception. It seems from these results as if the various components of a complete F0 contour produce the effect of a single "P-Centre" by directing perception to the correct part of the syllable. This suggests that the P-Centre itself does not have any reality, although the phenomenon of perceptual centring is real.

5 RESULTS

The ANSA system differs mainly from previous systems used to generate telephone numbers automatically by use of a sophisticated prosodic model to simulate natural pitch contours and rhythm. The hypothesis was that such a

system would produce speech that sounded more natural than similar earlier systems which did not have such a model.

Accordingly, tests were devised to assess the quality of the ANSA-produced speech in relation to both natural speech and the speech produced from the straight-line template system devised by Hall and Crombie [3]. Two sets of tests were used; firstly informal listening tests, and secondly visual tests to provide more concrete evidence.

5.1 LISTENING TESTS

In this test set, two types of test were used. The first was a straightforward listening test where recordings of telephone numbers spoken by a human operator were compared to their sythetic counterparts generated by the Hall and Crombie system and ANSA. The second was a test where the F0 parameter was separated from all other formant parameters and compared to the Lx waveform which had been simultaneously recorded on the database using a Laryngograph. This was done in order to force a comparison between pitch contours which as far as possible avoided all supraglottal parameters, such as voice quality, which might influence the decision.

The method in both cases was similar. Three versions of randomly selected telephone numbers were produced, one from the number recorded from a human speaker, the second from the Hall and Crombie system, and the third from ANSA.

5.1.1 LISTENING TEST RESULTS

In the case of time waveforms, all listeners identified the natural data instantly and agreed as to its superiority over either of the two synthesis systems. Next in order was the speech produced by the ANSA system. The Hall and Crombie system was generally agreed to be the least natural as regards F0, although the voice quality of both systems was judged to be similar, due to the fact that the same formant encoded templates were used for both sets of data.

The ANSA system was felt to perform particularly well in long number groups where the straight-line algorithm used by the Hall and Crombie system led to a very flat pitch contour. The difference between the systems was less marked in the shorter number groups, but was still evident.

The pitch tests produced more unexpected results. In this case, it was evident to all listeners which pitch was produced by the Hall and Crombie system, but it was harder to separate the natural Lx from the pitch produced by ANSA. When a forced choice was presented between ANSA and (natural) Lx, the Lx was detected with little better than chance frequency.

5.1.2 VISUAL TESTS

In this test set, F0 traces were obtained for natural and synthetic numbers. They were compared visually for discrepancies. Although this method should have produced concrete evidence of differences, it in fact produced a problem.

It was found from examining the two traces that generally there was good correspondence between the two, with minor variations. However, there were discrepancies, and it was not possible to say from visual examination alone whether these were due to faulty modelling of the intonation contour, or whether they represented legitimate variation in production.

Therefore, a further listening test was performed on the potentially wrong number group, with the following heuristic in mind:

IF, after listening to this number group, it is clear that it would never be produced in this way by a human speaker, THEN the contour is inappropriate.

This combined visual/aural approach highlighted some minor problems which were solved by adjusting prominences.

6 CONCLUSION

The ANSA software presented produces speech from formant encoded templates using a fully automatic synthetically generated intonation contour which accurately models most of the first-order phenomena found in natural pitch patterns.

From both sets of tests performed above, it was seen that this model gives ANSA a clear lead over earlier systems where a less sophisticated algorithm is used. This is particularly evident in longer series of digits, and when F0 is separated from all other parameters.

ANSA solves many of the problems encountered by previous systems by means of using an integrated phonological intonation model. This includes at least partial solving of the rather intractable problem of the P-Centre. The complete independence of the intonation model from the template data means that, in principle, any type of encoded speech may be used as input. In addition, ANSA may be generalised to deal with vocabularies other than telephone numbers.

Work currently in progress on improving the ANSA model involves collecting more data from speakers in order to model F0 more accurately.

REFERENCES

[1] **J.A.Waterworth (1983)** Effect of Intonation Form and Pause Durations of Automatic Telephone Number Announcements on Subjective Preference and Memory Performance. Applied Ergonomics, 14.1, pp. 39-42.

[2] **Olive, J.P. and Nakatani, L.H. (1973)** Rule-synthesis by Word Concatenation: a First Step. JASA Vol. 55, No. 3, March 1974, pp. 660 - 666.

[3] **Hall, M.C. and Crombie, J.R. (1988)** Digit Concatenation for Telephone Number Announcements Using Formant-Based Synthesis-by-Analysis. Proc. Speech '88, 7th FASE Symposium, Edinburgh. IOA, August 1988, Book 1, pp. 233-238.

[4] **Morton, J., Marcus, S. and Frankish, C. (1976)** Perceptual Centers (P-Centers), Psychological Review Vol. 83, No. 5, pp. 405-408.

[5] **Pierrehumbert, J. (1980)** *The Phonology and Phonetics of English Intonation*. PhD Dissertation, MIT.

[6] **Ladd, D.R. and Silverman, K.E.A. (1984)** Vowel Intrinsic Pitch of Vowels in Connected Speech. Phonetica, 41, pp. 31-40.

[7] **Lieberman, P. (1970)** A Study of Prosodic Features. Haskins Laboratories Status Report on Speech Research SR-23, pp. 179-208.

[8] **Taylor, H.C. (1933)** The Fundamental Pitch of English Vowels. Journal of Experimental Psychology, 16, pp. 565-582.