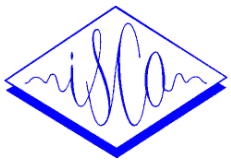

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

TEXT-TO-SPEECH SYNTHESIS USING A NATURAL VOICE SOURCE

Stephen D. Pearson, Hector R. Javkin
Speech Technology Laboratory
Division of Panasonic Technologies, Inc.
Santa Barbara, CA 93105, USA

Kenji Matsui
Central Research Laboratories,
Matsushita Electric Industrial Co., Ltd.
Osaka, Japan

Takahiro Kamai
Faculty of Engineering
Osaka University
Osaka, Japan

ABSTRACT

Our aim is to improve text-to-speech in its naturalness and its ability to model individual speakers. This paper describes various methods for using inverse-filtered waveforms from natural speech as a voice source in a text-to-speech system. One method uses a repeating loop, and controls pitch by interpolating samples in the waveform. Another method creates a source waveform of the desired pitch by concatenating single pulses from a collection of pulses. Listening tests were carried out to compare these methods with each other and with more traditional voice source generation techniques. The results indicate that these "natural glottal source" methods can substantially improve the quality of text-to-speech synthesis.

1. INTRODUCTION

Our principal goal is to improve naturalness in text-to-speech synthesis, while maintaining high intelligibility. In addition, we would like to have the means to switch automatically between several voice types, each closely matching the voice of one of our model speakers. Naturalness is strongly influenced by pitch contours, segmental timing, and formant values. One area where a great deal of improvement could be achieved is in the voice source. One idea, explored by Holmes[1], is to extract the source signal from natural speech by inverse filtering and use it directly in synthesis. This paper describes initial investigations into several methods based on this idea and applied to cascade formant synthesis based on a Klatt-type synthesizer[2].

In an effort to improve the naturalness of synthetic speech, we investigated the idea of using an excitation source extracted from natural speech. Such a source (one for each model speaker) inherently contains the proper spectral shape and time-domain characteristics. It is hoped that in this way some of the difficulties in creating natural-sounding speech with a parametric model can be side-stepped.

Holmes [1] and others [3] have used similar techniques in their parallel formant synthesizer. But in the cascade synthesizer, greater effort must be made to account for source dynamics, as opposed to the parallel synthesizer, which can, to some extent, account for source dynamics via formant amplitudes. Also, we are not aiming at perfect "copy" synthesis, but rather, practical source generation methods which can make a substantial improvement in our real-time, MITalk based text-to-speech system [4]. At present, we have implemented two relatively simple approaches to using natural speech for voicing source.

2. METHODS

In order to use sampled natural speech as excitation source in the context of a formant synthesizer, several steps must be taken. First, the sampled speech must be inverse filtered to remove the stronger, narrower resonances of the vocal tract. During synthesis, resonances appropriate for the current phonemic segment are regenerated in the formant filters, hence the original resonances should be absent from the source. Secondly, a method for pitch and amplitude control must be designed; we describe two such methods below. Thirdly, for high quality synthesis, a method is required to control other speech dynamics (for example, the change in broad spectral shape between /a/ and /m/).

We take a pragmatic approach: model whatever is practical in the synthesis filter, and account for the rest in the source. The inverse filter used in preparing the source can be, in some sense, the inverse of the structure of the synthesis filter. In this way, what we know how to model in the filter (formants, etc) is removed by inverse filtering, leaving what we don't know how to model for the source. Currently, our "synthesis filter" is a standard cascade of five 2nd order IIR filters to model formants, plus a pole/zero pair for nasal sounds and a parallel formant filter for fricatives [2] (figure 1). The inverse filtering was performed with the aid of an interactive computer system, with the filters adjusted by the investigator.

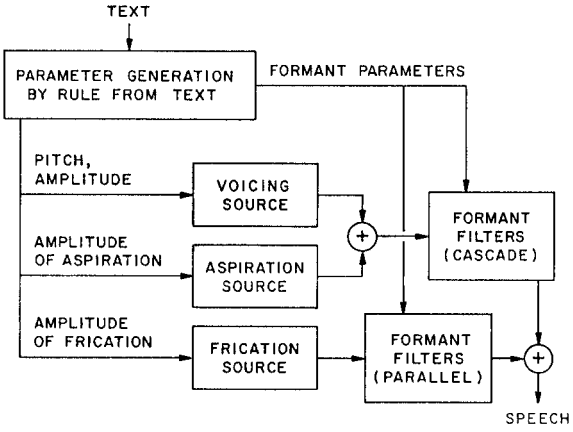

Figure 1: Synthesis Structure.

At present, we have developed some preliminary methods which focus on the essential requirement of prosody control. Amplitude modification is straightforward, and is handled by a multiplier. Pitch control, on the other hand, is a key

algorithmic consideration. There are two pitch control methods that we discuss in the following sections. These will be referred to as the "interpolated loop method", and the "pulse concatenation method".

Cycle to cycle variations in the source, jitter and shimmer, are important to naturalness. We have observed that a source which merely repeats a single inverse filtered glottal pulse results in a buzzy, machine like sound. Hence, we address this problem in our source methods. One solution is to use a sequence of glottal pulses which were contiguous in the original speech, thus inherent jitter and shimmer is preserved. A second solution is to introduce a randomness from pulse to pulse.

2.1 LOOP METHOD

The basic idea is to sample a real steady-state vowel, inverse filter this to remove vowel information, then use this data for the glottal source. The data is represented in a table which is used as a "loop". The table is read sequentially, and when the end of the table is reached, the next data is taken from the beginning of the table, and so on. To produce varying pitch, interpolation is performed within the table. In this way the table can be thought of as providing a continuous waveform which can be sampled with variable sample rate.

Several variations of the loop method are analyzed and compared for their effect on naturalness. The differences between these involve the order in which the loop data is used. We observe that these variations affect spectral characteristics, and jitter and shimmer.

Different loop methods:

- Single Period Loop (method A) A single period glottal pulse is repeated.
- Multiple Period Loop (method B) A natural glottal signal segment of a few pulse periods' length is repeated.
- Multiple Period Reversing Loop (method C) Pulses from a multiple period loop are selected in forward order, then reverse, and etc.
- Random Period Reversing Loop (method D) Method C is extended so that the direction reversal occurs at random points in the table.

In order to compare the spectral characteristics of the voicing sources derived from each loop method, we analyzed each source with constant pitch frequency. Figure 2 shows the spectrum of the original speech.

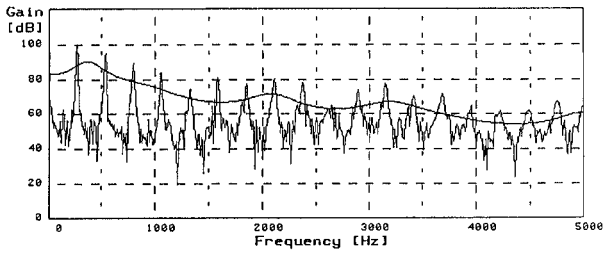

Figure 2. Spectrum of original speech sample.

Fig. 3 below shows the spectrum of each loop method. (2,048 point FFT, 12th order LPC)

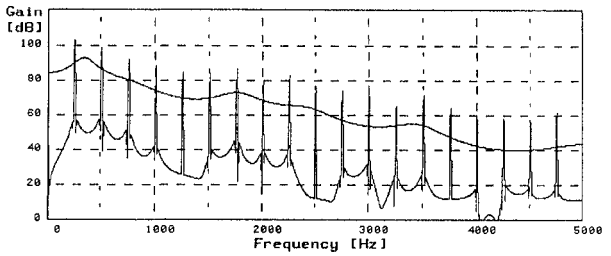

Figure 3a. The spectrum from method A. The spectrum consists of line spectra and displays almost no jitter or shimmer, because of the repetition of a single pulse.

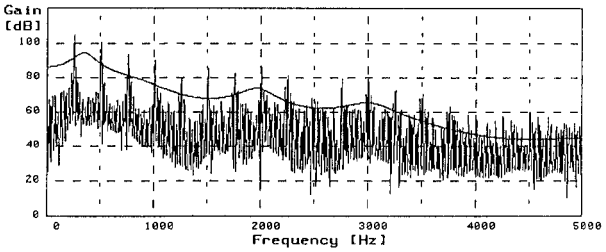

Figure 3b. The spectrum from method B. The original source has 10 pitch period pulses. The spectrum shows jitter or shimmer component, but the overall spectrum shape is different from the original natural source because of the spectral discontinuity between the final pulse and the first pulse of the glottal source.

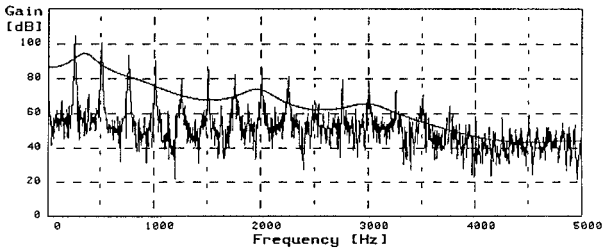

Figure 3c. Method C. Here the discontinuity problem is reduced, showing a more natural spectral shape than b.

In order to see the effect of method D, we looked at the frequency range between 0 and 1250 Hz, shown in figure 4, below.

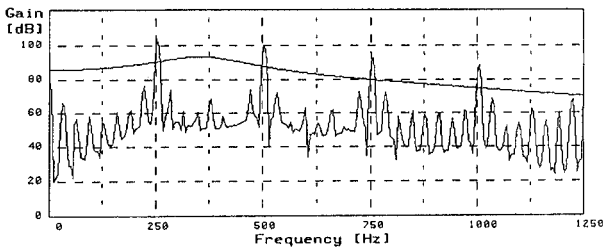

Figure 4a. Method C shows a periodic component, representing the loop repetition rate.

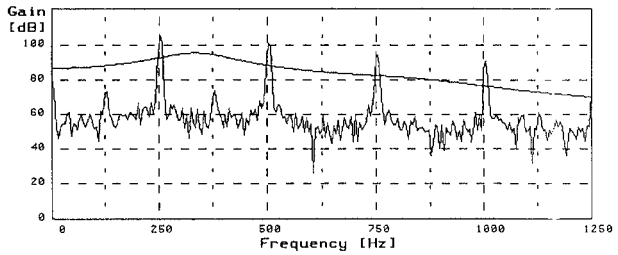

Figure 4b. Method D shows quite natural jitter or shimmer and no periodic component.

Experiment 1

A series of perceptual tests were conducted to determine (a) which of these methods and (b) what loop length yielded the highest ratings of naturalness. by human listeners [5]. Somewhat surprisingly, the results showed a preference for method A, the single pulse, which was judged to be very close to the original. Methods B, C and D showed very little difference. There was no significant difference between the results for 5, 10 or 20 cycles for any of the loop methods. We think this is because the loop source was not extracted from a very stable vowel. If the source for the loop has amplitude variation, any type of long loop based on it will also have an unnatural amplitude variation. As a next step, we have to study how to extract the best loop source from natural speech. Also we would like to continue this study using various words and sentences. During the perception experiments, subjects were asked to make their decision according to their preferences. We would like to try the same experiments having the subjects make the decision by similarity, not by preference.

2.2 CONCATENATED GLOTTAL PULSE METHOD

The concatenation method does not use interpolation to control pitch. Instead, in the simplest version, a library of glottal pulses, each with a different period, is stored in memory. The voice source generator produces a source wave train with a given input F0 contour by selecting and concatenating the appropriate glottal pulses. This avoids the problems that can occur with interpolation: distortion, spectral shift, and aliasing. The glottal pulses are derived from natural speech by first inverse filtering and then normalizing in amplitude. During synthesis, the amplitude envelope is imposed by multiplication with the amplitude control parameter.

The glottal pulses are spliced at a point during the closed phase. This minimizes discontinuities in the waveform or its derivatives. In addition, cross-fading (an overlap-add technique) is used near the point of concatenation to smoothly join the current and next glottal pulse. Results from implementing this method indicate that discontinuity problems are not critical. It is more important to obtain a good, compatible set of glottal pulses. We found that reducing the window of cross-fading, even down to one or two sample points, introduced little degradation.

The sampled speech data was recorded in a sound treated room using a Sony C-48 microphone, a sample rate of 10 KHz and a very sharp anti-aliasing filter at 5kHz. Approximately 300 glottal pulses from a single speaker with "reading style" voice were inverse filtered. About 50 of these were used to represent periods between 80 and 150 sample points. The 50 were selected from the higher amplitude vowel sounds and were also chosen for compatibility with each other. A few of the pulses are used to represent more than one input period. This is done by deleting some of the points from the ends of the waveform. This "waveform editing" is similar to many other approaches to time-domain prosodic modification (e.g. PSOLA [6] and microphonemic method [7]).

Since one glottal pulse is stored for each frequency, there will be slight random variations in the shape and amplitude from pulse to pulse. When these are concatenated, the variations have an effect similar to jitter and shimmer. However, if adjacent pulses vary too greatly from each other in spectral characteristics, a degradation of quality will result. Therefore, considerable effort went into selecting a compatible, smooth set of glottal pulses.

The glottal pulses are stored in the form of differentiated airflow. Therefore, no differentiation is necessary in the synthesizer to simulate lip radiation. An example pulse is shown in figure 5. The extreme minimum of the waveform is chosen to be the crucial time-point, T_{ref}, from which the splice points are relative. A smooth open-quotient (OQ) curve, as a function of period (in sample points) was selected to model empirical data.

$$OQ=\frac{120}{(120+period)}.$$

This is required in order to select a splice point which will occur in the closed phase of the glottal pulse. The splice point (calculated in terms of sample points) is given by:

$$T_{splice}=T_{peak}-OQ*period.$$

The chosen stored glottal pulse is then used as source, starting at T_{splice}, and for a number of points equal to the period.

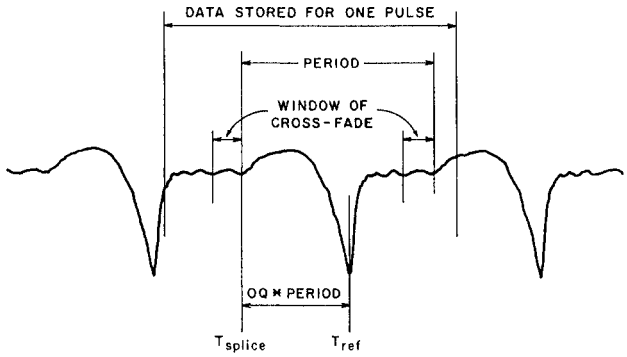

Figure 5. Example of Concatenation.

3. COMPARISON OF SOURCE METHODS

Experiment 2

An experiment was conducted to compare a sentence produced under 5 different conditions. Condition 1 consisted of natural speech spoken by a native speaker of Japanese digitized at 10 KHz with 16-bit resolution. The other 4 conditions were produced by extracting filter parameters from the natural speech, putting these parameters into our synthesizer and passing different source functions through it. Condition 2 used a glottal signal produced by a filtered pulse. Condition 3 used our parametric synthesizer [4] Condition 4 used a sampled voice source with the pitch controlled by the interpolation method described in section 2.1. and condition 5 used a sampled source with the pitch controlled by the concatenation method described in section 2.2. The stimuli consisted of comparison pairs of the speech produced under condition 1 fol-

lowed by speech produced under the five conditions (including condition 1). Subjects were asked to judge the similarity of the two members of the pair on a scale from 1 to 7. Eight repetitions were presented of each comparison, with the order reversed in half the presentations, for a total of 40 stimulus pairs. An additional set of 10 comparisons, labeled "practice", was included at the beginning of the experiment so that subjects could learn the range of difference of the stimuli. Subjects were told that they could alter their range of responses based on what they heard during this part, which was in fact not tabulated. The results of this experiment can be seen in table 1.

Table 1.

Condition	Description	Mean	Var.
1	natural speech	1.17	.20
2	parametric filtered pulse	5.45	1.61
3	parametric model	5.46	1.54
4	sampled loop method	4.05	.93
5	sampled concatenated	3.91	1.22

Natural speech (to itself) was significantly different than all of the other comparisons ($p < .005$). The two parametric methods were not significantly different from each other. The two sampled source methods were not significantly different from each other. Both parametric methods were significantly different from both sampled source methods ($p < .005$). Three classes thus emerge: natural speech, the two sampled source methods, and the two parametric methods. The two sampled source methods were judged significantly more similar to natural speech than the two parametric methods.

4. DISCUSSION

The methods described above are very successful at copy synthesis without hand tuning. An exception to this is the case of nasals or high vowels in the vicinity of nasals. For very high quality synthesis, the current source methods lack the flexibility to express a wide enough range of sound qualities. In this section we discuss some ideas for greater control of source dynamics.

Conceptually, the parametric source models are well suited to control source dynamics. There are difficulties, however, in finding the best set of parameters, and in generating these by rule. In addition, certain characteristics of the pulse appear not to be amenable to any sort of low-dimensional parametric model.

If the basis of source generation is inverse filtered speech from a variety of phonemic segments, a sufficient range of sound qualities can be created, and the problems associated with parameters are mostly avoided. The inherent correlations between source parameters are automatically represented in this data. The new problem is to select at each moment in synthesis the correct type of natural source data to use, and to pass smoothly from one type to the next.

We propose collecting a set of glottal pulses from a body of inverse filtered natural speech using a clustering technique. The set represents a particular speaker with a particular speaking style. The glottal pulses are like conventional CELP [8] vectors except that their lengths vary with the pitch period. Also, they contain greater short term correlation since our synthesis filter will not try to model the source. A "codebook" of glottal pulses with 250 to 1000 "vectors" is expected to be adequate and could include plosives and fricatives.

A distance metric will be needed to determine the closeness of two glottal pulses. The design of the glottal set ensures that more representatives are available for denser areas of this metric space applied to natural speech. An algorithm is needed to pick the best glottal pulse at any given time during synthesis. Having designed this pulse selection algorithm, most of the development work will be involved in creating the glottal pulse set and the associated structure for facilitating the selection process.

Within this method, the synthesis filter structure should account for formants and not much more. This follows since, in text-to-speech systems, additional structure in the synthesis filter will require new linguistic rules which are difficult to obtain. On the other hand, a more comprehensive model for the synthesis filter (with more rules) may allow the source to be represented by a smaller set of more uniform glottal pulses. A balance must be reached between these opposing directions. In addition, certain fixed properties could be factored out of the glottal pulse set: (1) A "voice quality" component of the synthesis filter which does not vary could be useful if it improved the time domain representation of the inverse filtered glottal pulses, and (2) gain can be factored out leaving only "shape".

There are several possible glottal pulse selection algorithms. One simply utilizes a table with relevant rule generated parameters as input and glottal pulse indices as values. But this table may turn out to be too large. A second method reduces this size requirement. Since, during synthesis, the next pulse to be used should be similar to the current one, only a small subset of the full inventory of pulses should be considered for the next choice. Hence, associated with each glottal pulse in the full set, we store a list of indices of other glottal pulses which are close enough to be considered for the next choice. In chosing the next pulse, the algorithm only examines pulses on that list to find the best candidate. The decision is based on information stored with each pulse. This may include period, amplitude, or possible phonemic context. This method can ensure that the next pulse will be similar to the current one, both in spectrum and in value at the point of concatenation.

5. CONCLUSION

Our work indicates that an extracted natural voice source can make a suitable voice source for synthesized speech. Although the methods we have utilized do not yet incorporate all the refinements we would like to include, the results of experiment 2 suggest that an extracted source can yield synthetic speech closer to a human model than our parametric models. To us, it seems likely that a more elaborated method for using such a source (in particular, methods that include a variety of source types) can improve further on these results and bring us closer to a natural-sounding text-to-speech system. Part of creating natural sounding speech is to create the voice, not of a "generic speaker", but of a particular individual. We believe that the methods we have outlined are especially suited to this task. The success of such a method can also help us to understand what elements are missing from parametric models. As a result, they can help focus on the elements which we do not yet understand in the voice source. We hope, therefore, that this work will not only help us produce improved text-to-speech, but also help us understand human speech production.

REFERENCES

[1] Holmes, J.N., IEEE Trans. Audio, Electroacoust., vol. AU-21, pp. 298-305, 1973.
[2] Klatt D.H., J. Acoust. Soc. Amer., vol. 67, pp. 971-995, Sept. 1980
[3] Howard D.M., et al, Proc. ICASSP, pp. 215-218, 1989.
[4] Javkin, H., et al, Proc. ICASSP, pp. 242-245, 1989.
[5] Kamai, T. et al, Fall meeting ASJ, 1-6-15, 1990.
[6] Hamon C., et al, Proc. ICASSP, pp. 238-241, 1989
[7] Lukaszewicz K., Proc. ICASSP, pp. 34.4.1-34.4.4, 1987.
[8] Schroeder M.R., et al., Proc. ICASSP, pp. 937-940, 1985.