Proceedings of the Institute of Acoustics

Naturalness in Synthetic Speech

Katherine Morton

Linguistics Department, University of Essex, Colchester

0. Introduction

The output from speech synthesis systems ideally should sound to the casual listener as though a recording of natural speech were being heard. At the moment, this can be done only when a good speech synthesizer is driven by parameter values directly extracted from actual human speech, as in copy synthesis or analyzed resynthesized speech. Text-to-speech systems produce the least natural sounding speech. These systems generate time-governed parameter values from tables of static values tied to individual speech 'segments', add a time dimension and link the segments by rule. The naturalness of different text-to-speech systems varies, but none has a quality which approaches copy synthesis or resynthesized speech.[1]

However, resynthesized speech is of limited use. The output is constrained by the original recorded human words and sentences which have been analysed. In contrast, text-to-speech synthesis is much more useful in principle, because of its versatility; new words and sentences can be created provided the combinations of segments required can be handled by the rules contained within the system. In comparing these two types of system, there is then a trade-off: the most general system sounds the least natural, and the least versatile system produces the most natural speech.

1. The research problem

If parametric speech synthesizers can be made to sound natural by controlling them with signals analyzed directly from human speech, then control signals generated in a text-to-speech system are inadequate or incorrect in some way. The differences between the two might be clearer if the output of both could be inspected side by side to determine the formal difference between them. The parameters of naturalness might then be determined by subtracting the one from the other. So far this has not been possible, and what exactly constitutes naturalness cannot yet be expressed in any way meaningful enough to enable improvement in text-to- speech systems.

It may be some time before researchers in speech will be able to specify segments sufficiently well for segment conjoining, and write satisfactory rules for conjoining them. In

[1] Resynthesized speech is the output of a process which automatically abstracts relevant acoustic characteristics of speech such as f0, formants and amplitudes, and formats these into a file suitable for input to a parametrically driven speech synthesizer.

addition, the underlying model of human speech which regards speech as a sequence of conjoined segments may be inadequate.

2. An alternative approach to synthesis

The approach to improved naturalness which I am describing in this paper attempts to extend the versatility and applicability of resynthesized parametrically analyzed speech. This approach does not directly tackle the definition of naturalness, but preserves its presence in real speech as much as possible.

What then would constitute extended versatility? Text-to-speech systems are versatile because they are based on modelling speech as varying linear combinations of a small number of discrete segments: these are usually allophone-sized units because this represents the practical minimum of units needed to generate all possible words in a language. The number is usually around 150. The general principle is that it is desirable to specify units that need minimum storage and can be rapidly retrieved. Each segment is represented by a single value for each synthesizer parameter, together with some additional information about its canonical duration and any special characteristics it has when conjoined with other segments.

However, it is possible to base a system on representations of larger speech units: syllables, words, phrases or sentences. The unit I have chosen is the word, each word specified as a complete file of synthesizer parameter values analyzed from neutrally spoken human speech. A value for each parameter is given for a series of time frames of 10ms, corresponding to the time frames of the JSRU synthesizer. Figure 1 (on the next page) shows part of the table entry for the parameter values of the conjoined digits *four* and *two* in the phrase *The number is three four two.*

Although resynthesized words will contain characteristics of naturalness, use of the system is limited to the number of words stored; to synthesize the entire language would require as many as 100,000 entries. Even so, the input text would probably require novel words that had not been recorded.

On a practical level, therefore the system described here could only be used where there is a limited vocabulary required and a restricted subject domain. For example, a small practical set could be the digits *zero* through *nine*. Such a system could define the domain of wrongly dialled telephone numbers, replacing the coded digitized recordings currently used by telephone companies. Extended somewhat, a digits only system could speak the time, give dates, bank balances, etc.

ms	FN	ALF	F1	A1	F2	A2	F3	A3	AHF	S	f0
10	250	48	550	14	800	40	2750	4	38	63	26
20	250	40	550	25	800	36	2700	1	42	63	26
30	250	9	325	9	1800	7	2800	1	5	1	25
.	250	5	325	8	1850	6	2850	1	6	1	25
.	250	8	300	9	1900	1	2900	2	7	1	25
	250	8	325	7	1900	17	2900	7	7	1	25
	250	8	350	1	1950	1	2950	1	9	1	25
	250	9	400	43	1950	28	2900	53	42	1	25
	250	5	450	45	1900	44	2250	56	54	1	25
	250	34	475	44	1700	38	2200	52	47	1	25
	250	36	525	46	1400	38	2150	51	54	1	25
	250	31	525	46	1350	33	2100	51	45	1	25
	250	51	425	47	1300	32	2050	56	59	63	25
	250	52	375	18	1350	30	2050	56	43	63	25

Fig.1 Conjoined parameter file of digit *four* followed by digit *two*. The first two rows are the final two frames of *four*, the next five frames are the stop phase of the [t], the next five frames the burst of the [t], and the last two frames the onset of [u] in *two*.

3. Difficulties in the approach

a. Outputting speech

Outputting speech by recalling the file of parameter values for a particular word in near real time presents no difficulties with small vocabularies. I have not addressed the problem of accessing large dictionaries.

b. The analysis

The analysis itself presents some difficulties. The analysis system I used was custom designed to extract the parameters for the JSRU synthesizer from speech sampled at 10kHz. These difficulties are usually experienced with such systems:

1. Formant tracking was variable during periods of very low amplitude; although this was offset to some extent by compression of the original digital recordings (using analogue recordings compression might have worsened the problem by bringing up the noise level in low amplitude signals).

2. Formant tracking was usually erratic during periods of unvoiced frication.

3. Extraction of formant amplitudes was sometimes inconsistent, especially when formants were close together in the frequency domain, and is a problem that often occurs with automatic parametric analysis. However, fundamental frequency tracking in general was fairly successful, except during unvoiced portions of the signal.

On the whole, the analysis system was able to provide useful parameter files for driving the JSRU synthesizer. Where there were errors, the files were post-edited by hand using the Laboratory interactive graphical editing software for synthesizer parameter files. Many errors can be located by listening to the output file; incorrect formant and amplitude values are signalled by 'clicks' and 'plops', and it is relatively simple to make corrections. Some obvious errors are unimportant: e.g. the specification of fundamental frequency during voiceless fricatives. In rare cases where no values were entered, numbers were interpolated between the last frame in which a valid value was supplied and the next frame containing a valid one.

c. The original human speech

The effort involved in dealing with resynthesized speech can be minimized by paying careful attention to the way in which the original recording is spoken. The criterion used here was to attempt to have the human speaker produce neutral speech. Cues to emotional content in the speech were avoided: no emphasis on specific words was intended, nor changes of speaking rate during or between items being recorded. The instruction given to the subject was to speak quite neutrally and to maintain a good rhythm of delivery.

The words recorded were not spoken in a frame, in order to eliminate possible coarticulatory tendencies at word boundaries. Consequently each word was spoken as though it were in itself a sentence, or the final word in a sentence. Thus characteristics of sentence-final words were retained in the analyzed speech. This would have presented no difficulties when the word was to be resynthesized in sentence-final position, but in other sentence positions such effects detracted from the naturalness of the final result.

Another major effect was that the speech rate slowed down from (and including) the final stressed syllable in the word. The effect applied to both vowel and consonantal segments. This effect does not always occur across speakers; therefore subjects being recorded should be tested before applying normalisation procedures.

4. Segmental conjoining

a. Trimming

Words uttered in isolation are typically longer in duration than when spoken within a sentence. Lengthening includes the final stressed syllable. Additionally, initial unvoiced sounds were somewhat longer than if they were to be used within a sentence. Quite detailed rules were devised to adjust word duration, but in this paper I shall outline some general rules which illustrate the approach taken in making the adjustments. The rules are presented here in prose. Two types of trimming were necessary: internal and external to the word.

Analyzed words are represented as two dimensional files in which rows represent synthesizer time frames and columns represent parameters (see Fig.1). Values for parameters are entered in the matrix cells for the given time frame. Files begin with the first frame of the word and end with the last frame. In the case of words beginning with a voiceless plosive the initial frames are those associated with the stop phase of the plosive, e.g. [t] as in *two*, although they may be silent when resynthesized. (I.e. all amplitude parameters set to one or near one, depending on the synthesizer.) Files of analyzed words need to be trimmed before they are used in resynthesis. The first rule applies at the beginning of words:

1. To shorten word-initial voiceless fricatives: if the file begins with a voiceless fricative (aperiodic excitation), trim all frames up to the first frame in which the amplitude of formant one reaches a predetermined threshold. This will vary among different systems, and should be set to indicate the frame which signals a change from background noise to the beginning of the speech. The amplitude level will be relative to the overall amplitude of the words in the recording session.

2. To shorten word-final voiced segments: if the file ends with a voiced sound (periodic excitation), trim from the end of the file those frames following the last frame in which the amplitude of formant one falls to a certain predetermined threshold as in 1.

3. To shorten word-final unvoiced segments: if the file ends with a voiceless sound (aperiodic excitation), trim from the end of the file those frames following the last frame in which the amplitude of formant three falls to a certain empirically determined threshold where A1 and A2 are high and A3 is low.

4. To trim a plosive burst: if the file ends with a plosive consonant (aperiodic excitation), trim from the end of the file those frames following the last frame in which the amplitudes of formants one, two and three are each simultaneously at a minimum (i.e. following the stop phase) (see Fig.2).

This rule removes the burst associated with plosive release, but retains the stop phase. Some English speakers delete the plosive release nearly always in sentence medial words, others do sometimes. The simplest rule is to always remove the release.

The above four rules are external trim rules. It may also be necessary to internally trim words spoken in isolation in order to shorten them.

ms	FN	ALF	F1	A1	F2	A2	F3	A3	AHF	S	f0
10	250	30	200	40	1950	27	2350	32	21	63	24
20	250	49	200	44	1950	36	2350	40	46	63	24
30	250	43	200	6	1950	27	2400	1	8	1	24
.	250	1	250	4	1950	21	2450	3	7	1	24
.	250	1	275	1	1950	26	2500	8	7	1	24
	250	1	300	1	1950	11	2550	4	7	1	24
	250	1	350	6	1950	3	2550	1	14	1	24
	250	1	425	2	1950	1	2500	1	7	1	24
	250	6	450	8	1950	12	2500	34	30	1	24
	250	1	450	4	1950	26	2450	33	48	1	24

Fig.2 Parameter file of the final frames of *eight*. The first two frames are the end of the vowel, the next six frames are the stop phase of [t], and the last two frames are the burst of [t].

5. If, after external trimming, the last frame of the file has aperiodic excitation, then count back the number of frames through the periodic excitation until the previous aperiodic excitation. This may be the frame of the end of the background noise. If the number of frames counted is greater than 30, then delete one third of the counted frames. If the number of frames counted is less than 31, then delete one fifth that number from the middle of the counted frames.

If the last frame of the file has periodic excitation count back the number of frames until the previous aperiodic excitation, and follow the same procedure for trimming as above.

b. Conjoining

Trimmed word files form the basis of the stored representations for resynthesis. The next part of the synthesis strategy is to assemble the words needed for conjoining (in the case of new sentences) or for inserting within an existing frame. Phonetic theory predicts that for a given word, attention will need to be paid to transitions from the end of the previous word and to the start of the next word.

The initial approach was to devise conjoining rules similar to those in the JSRU text-to-speech synthesis system (Holmes 1964). In practice, in many cases simply abutting the files was adequate. In the case of formant frequencies, if abutted parameters differed only by a few hertz (F1 – 50Hz, F2 – 500Hz, F3 – 700Hz) no interpolation was necessary. In most cases of formant amplitudes if the abutted parameters differed by less than 3dB then no interpolation was necessary. No appreciable gap in perceived speech was detectable in the final speech output. If the value difference was greater than these thresholds a normalization procedure needs to be applied to avoid a pumping effect. The simplest solution is to avoid the problem by recording the natural speech with minimal fluctuations of amplitude.

It seems that two mechanisms were responsible for the unexpected simplicity of the rules for conjoining. Firstly the listener's perceptual system smoothed over the join: listeners seem more tolerant of errors occurring in natural-sounding synthetic speech than in poor synthetic speech. Secondly the interpolators built into the synthesizer hardware used in this work will have smoothed the abrupt transitions (the implementation of the JSRU synthesizer used was from Loughborough Sound Images).

c. Prosodics

In conjoining resynthesized speech, two aspects of prosodics are important: rhythm and intonation. Rhythm presented little problem because the intrinsic rhythm of the original recordings (after trimming) formed the basis of the rhythm of the new phrase. Results were acceptable, producing a neutral speech output. This was judged satisfactory particularly as it provided the ideal starting point for an additional module to the system which acted upon a neutral input and, by rule, altered rhythm to overlay some emotional content on the speech output Adding emotional content to synthetic speech by rule is discussed in Morton *(forthcoming).*

This was also true in the case of inserting different words within a sentence frame. Since the sentence frame and word recordings were made using the same rhythm pattern, words fitted into appropriate slots adequately with respect to timing. The worst cases were encountered with polysyllabic words inserted into sentence frames which had gaps created by deleting monosyllabic words. Even here the effect was unusual, but not unnatural. Casual listening elicited comments like: 'a little strange..', 'a bit rough..', but not 'machine-like'.

In the case where phrases or sentences were to be created from stored words the input text was taken to the Laboratory intonation assignment algorithm. The f0 values are stripped from the file and a new f0 contour generated. The assignment procedure is a two-stage process involving assigning an abstract phonological contour (based on Pierrehumbert 1981), and generated with reference to a syntactic parse of the incoming sentence. In the second stage, the abstract contour is interpreted in terms of actual fundamental frequency values generated on a frame-by-frame basis according to the specified durations of the stored words (after Silverman 1987). The algorithm has been optimized such that when it fails, it produces a relatively flat neutral contour which may be judged unusual but not wrong. The results compare well with other systems.

A shorter procedure has been worked out for limited domains. In the case of the limited domain of digits, the intonation contour was specified by locating the final f0 value of the sentence frame phrase *The number is..*, adding 3, and entering this number into the first f0 frame for the first digit. The end of the frame for the first digit is located, 3 added to the value in the first frame, and values interpolated between the first and last frames. The same procedure is carried out for the second digit, except that 2 is added to both values. f0 for the third digit is calculated by interpolating between the end of the second digit and a 3 below the value at the end of the first digit. If a fourth number is needed, a value of 1 is added (between digits 2 and 3). If there are more than four digits a pause is inserted and the pattern changes. This procedure overcomes the effect of falling intonation on the isolated words in the original recording.

6. Summary

In this paper I have described an approach to the synthesis of novel utterances using limited vocabularies within restricted domains. The system uses the technique of resynthesis of parametrically analyzed speech. The units for building the synthesized output are words. Two methods are proposed: building entirely novel utterances by conjoining stored words, and using individual words to fill gaps in a parametrically analyzed sentence frame.

The trim rules used for isolating and normalizing words are generalizations which should work across different synthesis systems, but the threshold values used depend on the original human speech, the recording conditions, and the properties of the analyzer and synthesizer.

It is important to note here that strategies for synthesizing speech can be designed either to replicate as accurately as possible an actual human speech waveform, or they can be designed to provide a minimum sufficient set of cues for the listener to decode the speech output accurately and easily. In the work reported here I have been concerned with producing synthetic speech which satisfies those minimal criteria, yet at the same time preserves the naturalness of human speech.

7. References

J N Holmes, I G Mattingly and J N Shearme, 'Speech synthesis by rule', *Language and Speech* 7 p 127-143 (1964)

K Morton, 'Pragmatic phonetics', in *Advances in Speech, Hearing and Language Processing,* W A Ainsworth (ed.), JAI Press London *(forthcoming)*

J B Pierrehumbert, 'Synthesizing intonation', *JASA* 70 p 985-995 (1981)

K E A Silverman, *The Structure and Processing of Fundamental Frequency Contours,* PhD dissertation University of Cambridge (1987)