INVESTIGATING THE LIMITATIONS OF CONCATENATIVE SYNTHESIS

M. Edgington
Speech Technology Unit
Applied Research and Technology
BT Laboratories, IPSWICH IP5 3RE, UK
E-mail: mike.edgington@bt-sys.bt.co.uk

ABSTRACT

Concatenative text-to-speech (TTS) systems are now quite widespread through the availability of simple time-domain speech modification algorithms. Many of these systems produce intelligible speech with a higher degree of naturalness than that achieved by the previous generation of formant synthesis systems. This perceived improvement in quality has lead to the view in some circles that TTS is a solved problem, at least for many practical applications. Three experiments are reported in this paper, all performed with a concatenative TTS system. These experiments investigated aspects of the concatenative model by respectively addressing copy synthesis of emotional speech, modelling glottalisation, and the effect of speech database design on the quality of synthesised speech. This paper suggests that the lack of an explicit speech model in most concatenative synthesis strategies fundamentally limits the usefulness of many current systems to the relatively restricted task of 'neutral' spoken renderings of text, where deficiencies in other system components usually mask the limitations of the synthesis strategy itself.

1. INTRODUCTION

Over the past few years, the availability of simple time-domain speech modification algorithms, such as PSOLA [1], has lead to the widespread development of concatenative text-to-speech (TTS) synthesis systems. Some of these systems have been shown to produce good quality speech synthesis for many standard tasks [2]. The perceived quality of many systems has lead to an expectation of future performance which is often unrealistic, especially in respect of extensions to the simple task of converting plain text into speech.

In this paper some of the limitations imposed by the concatenative model are investigated by performing tasks which lie just beyond the normal requirements of TTS systems. In particular, these experiments attempted to address the limitations of the implicit speech model in most concatenative systems.

2. EMOTIONAL SYNTHESIS

The first experiment described in this paper explored the simulation of emotion using a time-domain synthesis model. This model allows explicit control of fundamental frequency, segmental duration and signal energy, which respectively relate to the perceived pitch, rhythm and loudness of the synthetic speech. In addition to these three parameters, most studies of emotional speech styles also emphasise the importance of voice quality through acoustic-phonetic features such as voice excitation mode, and pitch jitter [3]. Since the synthesis model does not allow direct control of these voice quality factors, the experiment attempted to discover how far emotional identity can be maintained without control of voice quality cues. A copy synthesis procedure was adopted to allow direct comparison between synthetic and natural speech, and to avoid the limitations of any specific rule system.

2.1. Emotional-Speech Database

The natural speech database comprised of four 'emotionally neutral' sentences, plus a five syllable reiterant phrase. Each sentence was produced by a trained actor in five emotional styles: anger, happiness, fear, sadness and boredom, along with a neutral rendition. This set of emotional styles included four 'primary' emotions plus boredom, which was expected to have distinct pitch and rhythmic characteristics. The recordings were made carefully in an acoustically treated studio, with both a field microphone and a laryngograph.

2.2. Re-synthesis

The prosodic cues of signal energy, syllabic duration and fundamental frequency were extracted from each sentence in the database. Signal energy and fundamental frequency were extracted automatically from the speech and laryngograph signals respectively, while the segmental durations were manually determined. These parameters were then imposed onto synthetic speech generated by a modern concatenative synthesiser [4], using a time-domain synthesis technique. Examples of angry resynthesised speech [A0743S01.WAV], default TTS speech [A0743S02.WAV], and original recording [A0743S03.WAV] are included in the CD-ROM version of this paper.

2.3. Subjective Assessment

A listening test was performed with 13 subjects, who had varying exposure to synthetic speech. The test was split into two sections, one consisting of the natural sentences and the other of the resynthesised sentences. (A pilot study suggested that mixing natural and synthetic stimuli

together distorted the results.) The order of the sections, and sentences within each section, was randomised for every subject. Each individual stimulus consisted of the same sentence presented twice, separated by a two second gap, and the subject was requested to choose which of the six possible emotional styles the sentence represented. In total 78 judgements were made for each emotional style.

2.4. Natural Speech

Table 1 shows the confusion matrix and recognition rates (RR) for natural speech. Although all emotions were recognised above chance (17% recognition rate), Anger and Happiness were very well recognised, while there was confusion between Fear and Sadness, and to a lesser extent with Neutral. Further analysis showed that certain specific sentences were responsible for many of the confusions, implying that the required emotion was not clearly produced in the recording. Removing the worst sentence for each of Fear, Sadness and Boredom, boosted their recognition rate to 56%, 68% and 85% respectively, with the average recognition increasing to 82.3%.

	N	A	H	F	S	B	RR (%)
N	69	-	-	-	5	4	88
A	-	78	-	-	-	-	100
H	1	-	76	1	-	-	97
F	8	1	4	41	23	1	53
S	12	-	2	13	46	5	59
B	10	3	-	-	3	62	79
						Average Recognition Rate (%)	79.3

Table 1 : Confusion Matrix for Natural Speech

These results are in broad agreement with recently published human performance on emotional speech databases, which gave average recognition rates of 75% [5] and 82% [6], each with databases of just four emotions.

2.4. Resynthesis

Table 2 shows the confusion matrix and recognition rates for the resynthesised sentences. The human recognition performance on the resynthesised sentences is clearly much worse than for natural speech; the error rate having almost trebled. As with the natural speech case, Fear and Sadness are the worst recognised emotions, but for resynthesis they are recognised no better than chance (17%). The high result for Neutral is caused by the subjects' tendency to overestimate the number of neutral utterances. Further analysis showed that the specific sentences which had the highest confusion in the natural sentences experiment were no more likely to produce confusion when resynthesised. No correlation was found between a subject's exposure to synthetic speech and their recognition rate.

	N	A	H	F	S	B	RR (%)
N	53	5	1	3	10	6	68
A	5	30	31	10	2	-	38
H	10	10	44	4	5	5	56
F	15	3	3	14	29	14	18
S	22	9	4	6	12	25	15
B	15	3	-	-	15	45	58
						Average Recognition Rate (%)	42.2

Table 2 : Confusion Matrix for Resynthesis

These results are broadly similar to those reported in [5], where the comparable recognition rate was around 30%.

2.6. Implications

This experiment confirms that imposing the prosodic parameters of fundamental frequency, segmental duration and energy from a natural speech utterance onto synthetic speech does not also result in the transfer of the perceived emotion. Excluding the confusable Fear and Sadness emotions, subjects were able to recognise emotions at 80-100% accuracy for natural speech, but only around 40-60% for resynthesised speech. This reduction in recognition rate has three possible causes:

- the chosen prosodic parameters do not carry sufficient information to clearly identify the emotion,
- lack of control of voice source characteristics is confounding perception of emotion, or,
- the speech modification process is introducing excessive distortion for the prosodic parameter range required.

All of these potential causes indicate that there is a fundamental limitation in the time-domain speech synthesis model which will prevent good quality emotional speech synthesis.

3. SYNTHESIZING GLOTTALISATION

The second experiment described in this paper investigated how the detailed control of fundamental frequency and local duration and amplitude could simulate certain segmental effects. The phenomenon of allophonic glottalisation was chosen as an appropriate segmental effect to study, based on the work in [7].

3.1. Allophonic Glottalisation

In English, the main uses of glottalisation are to reinforce word boundaries before vowel initial words, e.g. *Calgary airport*, and to resolve potential word boundary ambiguity in stop before sonorant contexts, e.g. *great eye* c.f. *grey tie*. Recent work [7] has described a strategy for synthesising this glottalisation by controlling the source parameters of a Klatt synthesiser. It was found necessary

to adjust the fundamental frequency (F_0), open quotient, spectral tilt and glottal flow rate in order to change the glottal pulse shape appropriately. In a time-domain synthesis model, there is no explicit control of the voice source characteristics, only the relatively gross signal characteristics can be changed. However, encouragingly it was reported in [7] that a lowering of F_0 in itself was a sufficient cue to glottalisation, although less natural than adjustment of all four parameters. Unfortunately, it was also reported in [7] that attempts to synthesise glottalisation using waveform concatenation had been unsuccessful.

3.2. Time-domain Synthesis of Glottalisation

The first approach to synthesising glottalisation, the basic model, was simply to apply the F_0 reduction suggested in [7], i.e. lower F_0 to 35Hz (50Hz for female speech) for a duration of 100ms. Informal listening tests showed that this simple approach produced a sound that was variously described as 'purring', or 'a bit like creaky voice', but was not very similar to glottalisation. Adjustment of the target F_0 to other frequencies did not improve the perception. A side effect of reducing F_0 for 100ms was to mask part or most of the following sonorant, reducing the intelligibility of the synthetic speech.

In order to maintain intelligibility, the basic F_0 reduction method was enhanced to first extend the duration of the sound preceding the glottalisation by 100ms, then impose the standard F_0 reduction. Although intelligibility was satisfactorily maintained with this method, the unnatural sound of the glottalisation still remained.

3.3. Analysis of Glottalisation Database

In order to investigate glottalisation in more detail, a small database of one male and one female speaker was created. The database consisted of 12 complete phrases with vowel-vowel contexts across word boundaries, and 10 pairs of phrases which exhibited potential word boundary ambiguity. Each of the first 12 phrases were recorded twice by each speaker, once with and once without glottalisation. The ambiguous phrase pairs were also recorded with and without glottalisation. The speakers were recorded using a microphone and laryngograph, so that inappropriate glottalisation could be identified during recording, allowing correct repetition of the phrase.

Analysis of the speech database showed that there were three consistent features found around the glottalisation:

- there was a decrease in F_0 during the approach to glottalisation,
- the speech signal was significantly attenuated (usually to silence) during the glottalised portion
- there was no consistent pattern to the F_0 contour when voicing resumed

The first set of vowel-vowel phrases were used to measure the extent of the observed features. The median results were:

- minimum F_0 value prior to glottalisation: **70Hz**
- duration of approach: **20ms**
- duration of silence: **69ms** (male) **57ms** (female)

The first two values were consistent between the two speakers. The difference in the silence duration might be due to the difference in speaking rates: the female speaker averaged 246 words per minute, and the male 222 words per minute. Analysis of the ambiguous phrases was more difficult due to preceding stops, but was broadly in agreement with the above figures. Since considerable care was taken in the recording process, it is likely that these figures represent the characteristics of deliberate glottalisation, not spontaneous occurrences, but this is entirely appropriate for use in a TTS system.

3.4. Improved Glottalisation Synthesis

Analysis of the glottalisation database suggested a better procedure for time-domain synthesis of glottalisation. The fundamental frequency of the speech signal is reduced to 70Hz in the 20ms preceding glottalisation, and 70ms of silence is then introduced. To avoid any signal discontinuities, the signal amplitude is linearly reduced in the 3ms before the silence, and linearly increased in the 3ms following the silence.

3.5. Evaluation

A listening test was conducted to rank the three synthesis methods, by performing a set of pair-wise comparisons against a reference (often called ABX tests). Each stimulus consisted of the same unglottalised phrase modified by two of the methods (A & B), then the natural glottalised phrase as a reference (X). Subjects were asked to judge which of A & B *is most similar to* the reference X. Three of the vowel-vowel phrases were chosen for the test, each phrase occurring in six stimuli (each pairing of the three methods, in both orders), to 14 subjects. Student's t-distribution was used to analyse the data, and the results are summarised below:

- Improved model was more natural than the basic model at 99% significance level
- Improved model was more natural than the basic + duration model at 99% significance level
- Basic + duration model was more natural than the basic model at 90% significance level

A further listening test was conducted to assess the effectiveness of the improved model. Four of the

ambiguous phrases were chosen from the database. Subjects were presented with one of three versions:

1. the original glottalised recording,
2. the original unglottalised recording,
3. the unglottalised recording modified by the improved glottalisation synthesis method.

The subjects were asked to judge which interpretation of the phrase they heard. The correct interpretation rates are presented in Table 3. Although the number of phrases is small, these results does provide evidence that the modified glottalisation synthesis model is generally successful.

phrase	glottal -ised	unglottal- ised	synthetic
sea liner - seat liner	100%	100%	93%
grey tie - great eye	89%	100%	93%
see Mabel - seem able	71%	100%	96%
heavy yoke - heavy oak	96%	100%	86%

Table 3 : Ambiguous Phrase Correct Interpretations

4. NATURE OF THE SPEECH DATABASE

The last set of experiments set out to investigate how the characteristics of the speech database used for concatenation affects the quality of the synthetic speech. To demonstrate these effects, a set of words and phrases were synthesised using three different speech databases:

1. a diphone database
2. a demi-syllable database
3. phonetically-balanced continuous speech database.

The selection of speech units for the second and third databases used a non-uniform unit selection algorithm which incorporates a structured phonological model of the database. All other parts of the synthesis procedure remained constant.

Informal listening tests indicate that the perceptual quality of the synthetic speech is very strongly determined by the nature of the original speech database. For example, single words synthesised from the demi-syllable database [A0743S04.WAV] were more articulated than those from the continuous speech database [A0743S05.WAV], and were accordingly judged to be clearer. In contrast, sentences were perceived as being more natural when they were synthesised from the continual speech database, but over-articulated when using the other two databases.

5. CONCLUSION

This paper considered some of the limitations of the standard concatenative synthesis model by investigating the use of a typical TTS speech model for tasks which lie just outside the domain of current TTS. Work on synthesis of emotional speech indicates that time-domain models, at least, are insufficient for this task, and it is currently unclear whether basic prosodic parameters can even encode enough information to reliably synthesise emotion. It is therefore likely that a far more sophisticated speech model is necessary for emotional speech synthesis. However, experiments have shown that a simple time-domain model can be successful in the synthesis of glottalisation, which is a segmental rather than prosodic effect.

The evidence gathered so far from use of different speech databases, indicates that the nature of the database is a critical factor in the quality of the synthetic speech. This implies that truly flexible concatenative TTS systems should incorporate a speech model that is able to modify more than just the basic prosodic characteristics of the speech signal. Alternatively the TTS system should use a sufficiently large database to include all required styles of speech, and a speech unit selection procedure which incorporates a speech model which has the sophistication to determine the speech style of sections of the database. In any case a limiting factor in the development and widespread acceptance of concatenative TTS systems is the simplicity of most current speech modification models.

6. ACKNOWLEDGEMENTS

The author would like to thank Benjamin Elson and Alan Gee, both of the University of York, for their efforts and enthusiasm in implementing some of these experiments.

7. REFERENCES

[1] E.Moulines & F.Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communications*, Vol. 9, pp 453-467, Dec. 1990.

[2] R.D.Johnston, "Beyond intelligibility - the performance of text-to-speech synthesisers", *BT Technology Journal*, Vol 14, pp. 100-111, Jan. 1996.

[3] I.R.Murray & J.L.Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *J. Acoust. Soc. Am.*, Vol 93, pp. 1097-1108, Feb. 1993.

[4] J.H.Page & A.P.Breen, "The Laureate text-to-speech system", *BT Technology Journal*, Vol 14, pp. 57-67, Jan. 1996.

[5] B.Heuft, T.Portele & M.Rauth "Emotions in time-domain synthesis", Proc. ICSLP'96, pp 1974-1977, Philadelphia, USA, 1996.

[6] F.Dellaert, T.Polzin & A.Waibel, "Recognizing Emotion in Speech", Proc. ICSLP'96, pp 1970-1973, Philadelphia, USA, 1996.

[7] J.Pierrehumbert & S.Frisch, "Synthesizing Allophony Glottalization", in *Progress in Speech Synthesis,* eds. Van Santen et. al., Springer-Verlag, New York, 1996.