# Speech Prosodics for Synthesis – Perspectives

**Mark Tatham**
**Katherine Morton**

Reproduced from: Tatham, M., and Morton, K. (2000), Speech prosodics for synthesis – perspectives. In Botinis, A. (ed.) *Fonetik 2000*. Skövde: Högskolan Skövde. 133-136



---

## Abstract

Speech synthesis systems still fail in producing acceptable prosodies. We are developing a research strategy designed to de-focus attention on the objective acoustic accuracy of synthetic speech in favour of enhancing the speech to optimize a listener's ability to repair 'damaged' signals. To do this we need to know more about how listeners repair errors and how we might trigger the repair processes: we believe that there is much to be gained from this approach to improved speech synthesis.

## Introduction

Despite improvements in overall voice quality in modern synthetic, prosodic rendering still fails to reach a quality or naturalness acceptable for the wide deployment of synthesis in information retrieval systems (Dutoit, 1977; Tatham, 1993). In typical text-to-speech synthesis the most heavily computed area is the prosodies. This leads us to believe that current models on which the computation is based fail us.

A survey of six synthesisers: the original 1965 Holmes parallel formant device, followed by the current versions of Edinburgh Festival, OGI Festival (an American version of the Edinburgh original), AT&T, a bad and therefore unnamed synthesiser, and the SPRUCE system (Lewis & Tatham, 1999; Tatham, 1993), reveals that in all cases it is clear that whatever the shortcomings of the segmental rendering – the way in which the individual sounds are produced and coarticulated – the major fault lies with their rhythm, stress and intonation.

## Perspective

Building on the idea that the goal of inter-human communication to copy sender plans to receiver percepts we discuss whether a model of prosodies production could benefit from an inclusion of a sub-model of prosodies perception. Put simply, we discuss how a greater understanding of prosodies perception might influence our approach to modelling production. Understanding the extent to which repair of signal degradation figures in perception enables us to re-appraise our approach to just how *accurate* production modelling should be. If the goal of speech production is to create a percept copy in the perceiver how might a perceiver's repair strategies be exploited to offset errors in production? This general principle has been around for a couple of decades in linguistic phonetic theory – it is worth dusting off again for possible inclusion in synthesis strategy.

Speech acoustic signals seem to be a degraded representation of a speaker's plan. In linguistics terms the plan is developed in the cognitive phonological stage of speech production, to be handed over to the phonetics stages for implementation. It is as we pass from an abstract cognitive representation to a physical representation that problems of degradation begin. Speech motor control and production in general are far from perfect – errors are introduced during all stages of the physical processing (Pierrehumbert, 1990). Examples are to be found in coarticulatory phenomena. Furthermore signals received further

damage from the environment: ambient noise or transmission deficiencies further degrade the signal.

The task of the perceiver is to 'recover' the original plan from the degraded signal (Bregman, 1990). There are two ways of looking at this: either the signal still contains sufficient of the plan to enable recovery, or the perceiver sets about an active process or reconstruction or repair. We favour the latter model, believing that even completely lost aspects of the plan can be reconstructed to create a percept which is an adequate copy of the speaker's plan. When this happens recognition is complete. Our perspective is therefore from cognitive science. We believe that speaker and listener collaborate to ensure good transmission of thoughts and ideas between them (Fougeron & Keating, 1977). That collaboration is active, we feel, and ready to be incorporated in speech synthesis systems.

The basis of our approach is to put forward the idea that if we can understand more of the perceptual strategies involved in repair we can make changes to our synthesised signal which can lead directly to optimising perception. When a listener declares a stretch of synthetic speech to be bad, they may simply be reporting inability to carry out complete repair. If we cannot yet make the signal intrinsically better, we may at least be able to do (unnatural) things to it which will assist the listener. Our strategy an explicit model of repair in perception – something we believe to be currently lacking.

## Trials and examples

The basis of prosodic modelling in linguistics is the syllable – a notoriously elusive concept rivalling only the phoneme for its ephemeral convolutions. But 'syllables' are useful synthesis building blocks, or they ought to be when well defined. We discuss definitions, offer our own, and show how we use these in manipulating (Tatham, 1995) high-level areas of our synthesis (its phonology, in linguists' terms) and in low-level areas of our synthesis (its phonetics). We examine in particular notions of rhythmic alignment in respect of 'telescoped' syllables, and conclude with a discussion of the phonological phenomenon known as 'ambisyllabicity', and how this might descend to a phonetic level when recorded syllables are chained in waveform synthesis.

By way of illustration we use examples based on comparison between human and the SPRUCE rendering of the same complex sentence. We try to make the modelling problem explicit by presenting graphs for each of the acoustic correlates of prosodic phenomena:

- rhythm – segment and overall timing,
- stress – segment intensity, segment timing and local manipulation of fundamental frequency,
- intonation – global manipulation of fundamental frequency across the domain of the utterance.

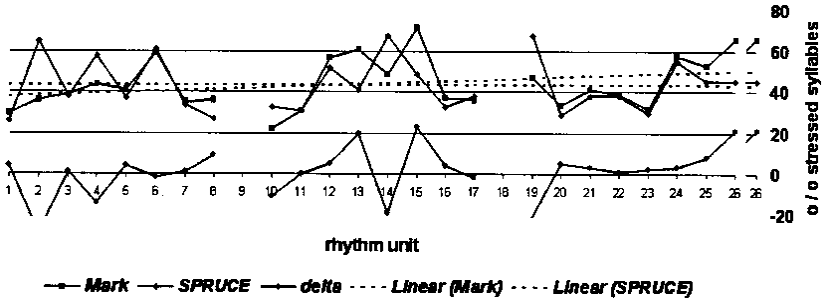


Fig. 1 Durations between the onsets of stressed syllables – indicating the timing of rhythm units. Note the decrease in syllable timing for the human talker toward the end of the sentence. The graph separates the three phrases within the sentence: "Over most of the country there will be lengthy spells of clear weather, but in many areas of Wales and across southern and eastern England there will be mist with some rather thicker patches of fog in places."

In each case the linguists' terms reside in the phonology, are cognitive in origin and treated as abstractions in linguistic theory. They do, however, have acoustic correlates, though these may not exist on a one-to-one basis. We must model the acoustic correlates and show how they trigger the perceptual recovery of the original plan.

Our data show, for example, that toward the end of each phrase in the test sentence there is a gradual increase of segmental duration, as well as the duration of the intonational unit 'foot' in the human version, whereas the synthetic version fails to achieve this adequately – thus failing to cue, for example, turn taking if this sentence were part of a dialogue. A linear regression plot of the data clearly shows the problem when we graph durations between stressed syllables [4]. Graphs of intensity show also that the human talker gradually reduces relative intensity as phrases proceed, finally dropping intensity significantly toward the end of the sentence. The synthesis does not do this.

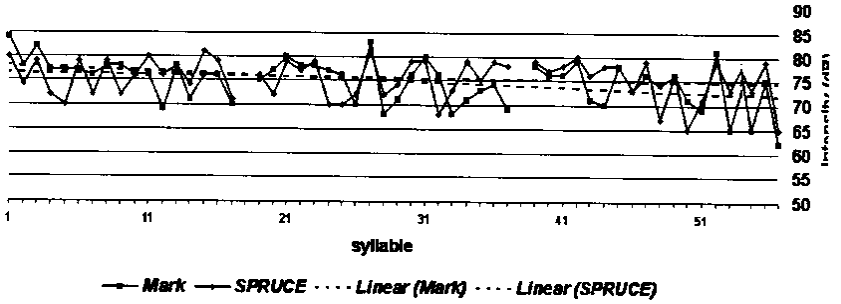


Fig. 2 Intensities of all syllables throughout the test sentence. Note the decreasing intensity for the human talker in each of the three phrases (particularly the final one) as the end of the phrase is reached. The sentence is the same as in Fig. 1.

When we come to intonation we find the worst examples of flawed prosodics in the synthetic speech. The model used in SPRUCE is not entirely unique, and has been described elsewhere in some detail. We believe it is not too bad, even in dealing with long sentences of the kind used to illustrate this presentation. However, just how bad is it? We bring the problem home by using our PSOLA research tool to take the human voice and give it SPRUCE intonation – the result is obviously 'wrong', but casual observation might feel that it is more repairable than when used for the synthetic speech. Instantly, though, with perfect segmental rendering we can use this tool to focus on just what the errors are. We are systematically introducing generated errors into the otherwise 'perfect' human version to judge listener reaction.

To emphasise the point further we present each of the three phrases of the test sentence separately in synthesised version – but with the human intonation directly transferred to the signal. This is achieved hand copying, but is extremely revealing when used systematically. We can gradually expose the listener to improvements or degradations – and judge their reactions (Morton, Tatham & Lewis, 1999).

## Conclusion

Our paper is about syllable-based synthesis and prosodics. It considers perspective carefully and suggests that a particular way of looking at the overall communication system (speaker and perceiver together) prompts a special approach to speech production. We use this idea to give examples of a synthesis strategy in the prosodics area resulting in demonstrably improved naturalness.

Our perspective in approaching speech synthesis is to set as goal the creation of a good percept in the mind of the listener – a percept which truly represents the speaker's plan and which minimises the work the listener must do to achieve the right repairs to our relatively poor synthetic signal. By systematically observing listeners' reactions to progressively improved or degraded signals we are gradually building a model of the repair processes.

The idea is to introduce into our synthetic speech special repair-oriented cues which, we hope, will cause the listener to report an improvement in the quality of the signal, *although the signal itself is no closer to a human signal than it was before*. What we are doing is creating synthetic speech *to be perceived*, not synthetic speech which is to be tested against a natural signal. While we wait for a perfect production model geared to the needs of synthesis, we feel that we can make significant progress with this alternative approach.

## REFERENCES


Bregman A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. Cambridge MA: MIT Press

Dutoit, T., 1977. An Introduction to Text-to-Speech Synthesis. Dordrecht: Kluwer Academic

Fougeron C, and Keating P., 1977. Articulatory strengthening at edges of prosodic domains. JASA 106:6, pp. 3728-3740

Lehiste I., 1977. Isochrony reconsidered. Journal of Phonetics 5, pp. 253-263

Lewis E., and Tatham M., 1999. Word and syllable concatenation in text-to-speech synthesis. Proceedings of Eurospeech '99 (G. Gordos ed.). Budapest: European Speech Communication Association, CD-rom

Morton, K., Tatham, M. and Lewis E., 1999. A new intonation model for text-to-speech synthesis. Proc. ICPhS (Ohala, J. et al. eds) University of California at Berkeley, CD-rom

Pierrehumbert J.P., 1990. Phonological and phonetic representation. Journal of Phonetics 18, pp. 375-394

Tatham M., 1993. Voice output for human-machine interaction. Interactive Speech Technology (Baber, C. and Noyes, J. eds). London: Taylor and Francis, pp. 25-35

Tatham M., 1995. The supervision of speech production. Levels in Speech Communication (Sorin, C. et al. eds.). Amsterdam: Elsevier, pp. 115-25