SPEECH SYNTHESIS, SPEECH SIMULATION AND SPEECH SCIENCE

Mark Huckvale

Department of Phonetics and Linguistics
University College London, London, U.K.
M.Huckvale@ucl.ac.uk

ABSTRACT

Speech synthesis research has been transformed in recent years through the exploitation of speech corpora – both for statistical modelling and as a source of signals for concatenative synthesis. This revolution in methodology and the new techniques it brings calls into question the received wisdom that better computer voice output will come from a better understanding of how humans produce speech. This paper discusses the relationship between this new technology of simulated speech and the traditional aims of speech science. The paper suggests that the goal of speech simulation frees engineers from inadequate linguistic and physiological descriptions of speech. But at the same time, it leaves speech scientists free to return to their proper goal of building a computational model of human speech production.

1. INTRODUCTION

Statistical modelling and concatenative signal generation techniques have come to dominate the field of speech synthesis over the last decade. Contemporary models of timing (e.g. [17]), of intonation (e.g. [6]), prosodic phrasing (e.g. [4]), or pronunciation (e.g. [7]) are based on statistical analysis of labelled corpora. In signal generation meanwhile, vocal tract models have been replaced by first diphones (e.g. [9]), then poly-phones (e.g. [1]) then arbitrary length segments (e.g. [16]), such that the dominant method for making new speech signals in 2002 is to build them from bits of old speech signals. These changes and the new techniques that drive them seem to me to question the relationship between speech technology and speech science. In particular the shift away from knowledge-based rules and vocal-tract processing challenges a previously accepted principle: that *better speech technology will come from better speech science*. Or that better voice output will come from a better understanding of how humans produce speech. It is not just that concatenative synthesis and statistical linguistic analysis are ways to cover up our lack of knowledge about how human speech is produced, but that current research into improving speech synthesis involves methods (like unit selection) that do not contribute to an understanding of human production at all. In our efforts to make the best sounding speech with these techniques we seem to have moved away from the scientific study of speech.

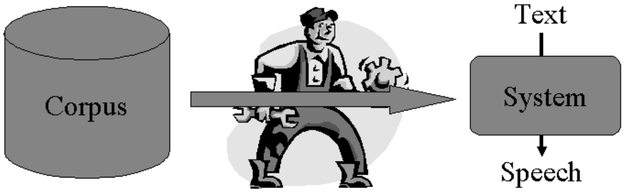

Figure 1 Old style speech synthesis research

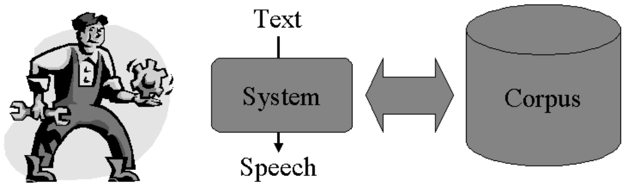

Figure 2 New style speech synthesis research

To understand this let's see how speech synthesis research has changed. I believe the mental picture many speech scientists had of speech synthesis research before the current era was something like Fig.1 – the researcher delved into a corpus of speech recordings looking for recurring patterns, and then encoded these as 'rules' which could be incorporated into a TTS system, often in the form of hard-coded symbol manipulation. These rules were divined by the researcher and encoded in the system and so it was possible to say that knowledge about speech was being discovered and exploited. It appeared that we were doing speech science at the same time as building speech technology. At the same time as our system got better, we also had a formal description of why. In contrast, the current approach is more like Fig.2 - where the researcher stands to one side and lets the system delve into the corpus of recordings itself. The scientific input of the researcher is largely limited to labelling the corpus – using linguistic knowledge to demonstrate which sections of the recordings are linguistically "equivalent". The system is provided with some machine learning algorithms, such as CART [11], analogical reasoning [3], or statistical signal processing [15] and also some objective metric to maximise. It is then left on its own to decide how best to produce sounds that are similar to the recordings found in the corpus. The system may find and exploit linguistic patterns or systematic variation in phonetic form, but that knowledge is not explicit, is not coded as rules, and does not contribute to "understanding" in scientific terms. For example, a CART tree used for a duration model has typically thousands of nodes. Not only is it impossible to comprehend the tree as an entity, it also has too many free parameters to constitute a "theory". Indeed the power of a CART model is that it makes very few assumptions about the statistical properties of the data.

But even if corpus methods are not contributing to speech science, there is no disputing the fact that *on the average* the quality of generated speech has improved markedly through their use. We seem to have better systems through the application of less explicit knowledge. If there are criticisms to be made, then they need to be made of speech science: why has it not delivered the theories and knowledge necessary for rule-based synthesis? Is it really then case that corpus methods are just a stopgap measure, as Shadle and Damper [12] suggest? What are the limitations of these techniques? Indeed what are the new goals of speech synthesis technology?

In this paper I will look at three main issues: (i) what is a suitable goal for speech synthesis technology, (ii) how does this goal set a new agenda for research, and (iii) what future is there for a science of human speech production? After reading this, I hope you will be able to say where you stand on the debate between technology and science.

2. REPLICATION, UNDERSTANDING OR SIMULATION?

The Stanley Kubrick / Arthur C Clarke film *2001 A Space Odyssey* contained the most famous talking computer, HAL [14]. I think that many researchers in speech see the *replication* of human speech the goal of speech synthesis research. The aim is to create a system like HAL that uses speech as a human does: as a means of intentional communication, as a means of expressing understanding, desires and emotions. But although it is possible to say that the only difference between the form of HAL's speech and the speech of a modern TTS system is one of natural prosody, there are enormous differences between HAL's use of speech and that of a modern TTS system. We are not in any sense close to *replicating* how humans or HAL use speech because we are nowhere close to building a system with intentionality, with the need to communicate. The lack of expression or interest or emotion in the speech of current systems is due to the fact that the systems don't actually understand what they are saying, see no purpose to the communication, nor actually have any desires of their own. It makes no sense to add "expressiveness" or "emotion" to voice output from a mindless system.

But if the goal of speech synthesis is not to replicate a human talker, is it to *understand* how humans talk? In other words to build a computational model of human speech production that actually talks. Are corpus-based methods just a deviation from the one true path to the goal of a theory of speech? This might be true if the methods they replaced were also on that path. But right from Jim Flanagan's denotation of a formant synthesizer as "terminal analogue" device [5], synthesis research has used simplified models of the process of speaking to get the job done. What components of any text-to-speech system operate in any way like a cognitive model? Not message-generation, nor pronunciation, nor planning, nor articulation, nor sound generation. Even admirable attempts to make speech synthesis more "articulatory" like HLSyn [13] serve just to patch up inadequacies of a fundamentally non-human production system. Of course there is excellent research into "concept-to-speech" which gives systems more information about how to generate suitable prosody [8], or research into how non-linear phonological descriptions might model systematic phonetic variation in context [10]. But these are ideas operating within a TTS framework, not a cognitive one; the hope is to make more convincing synthetic speech not understand how humans do it.

If the goal of speech synthesis is neither replication nor understanding, then I suggest it must be *simulation*. We ask our system to take a linguistic input and produce a noise that is similar to the noise a human being would make for the same request. Or more precisely: we ask the system to produce a noise that a human listener believes is similar to the noise a human being would make for the same request. We record our speaker in the studio reading a sentence, and we want our system to produce a noise from that sentence such that it sounds to us linguistically equivalent. In simulation, there are no constraints on how the system achieves its goal. There are no requirements to use a vocal tract to generate the sound, to model the motor control of articulators, to use accepted phonological analysis, to know what the words mean individually, to know what is implied by the text, or to know how the style of speaking relates to the communicative context. In simulation, knowledge of human speech production is optional: it may help, such as the use of phonological units for labelling substitutable regions of signals, but it may not, such as enforcing a source-filter separation in prosody manipulation. Importantly, the techniques that work well in simulation may not be relevant to human production: unit selection, audio morphing, spectral smoothing, or overlap-add.

Thus I see the goal of current speech synthesis research is to make better simulations: fake speech that fools more of us more of the time. This is a significant shift away from older ideas of replication and understanding. One might even argue that the term speech synthesis itself is outdated. Speech synthesis implies that new utterances are built up from simple components. But simulation can encompass plain audio recordings, or slot-and-filler announcement systems. A speech simulation system need not start with elementary building blocks; it might just build new utterances from old ones.

3. ROUTES TO BETTER SIMULATIONS

Viewing the task of speech synthesis research as that of building better simulations with no constraint to model human production is actually liberating. It is no longer necessary for a process to be natural (such as concatenation) nor produce ultimate performance (such as prosody manipulation). We don't even need to know whether the process is always possible (such as voice conversion) for it to be useful in some circumstances. Here are some suggestions for research topics in the science of speech simulation itself:

More natural corpora: current corpora are collected from read speech, where the speaker maps text to speech, so it is not surprising that our systems sound like simulations of reading. Speech signals collected in more natural communication contexts or in different styles will produce speech simulated in different styles. To simulate a weather forecaster or a telephone operator there will be nothing better than a corpus of real forecasts or real operator conversations.

Higher-level linguistic descriptions: it would be no use just labelling these more natural corpora with just orthographic transcription. These styles are not just read text and deserved to be labelled with richer transcriptions: with details of the dialogue acts or of rhetorical structure, for example. This will have to be in concert with information systems that can deliver higher-level descriptions of what they want to say [8].

Richer phonetic labelling: phonological and phonetic labelling is used in unit selection to identify regions that can be substituted for one another. But the regions that can be substituted phonologically (in terms of contrast) are not necessarily the regions that are phonetically equivalent (in terms of articulatory form). Trivially, /t/ is distinctly different in *tie*, *try* and *sty*. But in simulation, why be limited by phonology? Shouldn't we be using a phonetic description based on what regions can and what cannot be substituted for each other? A phonetic description acquired by data-driven methods might even lead to superior phonetic labelling with less manual intervention.

Machine learning: in simulation there is no need for a debate about the structure of a statistical model. Models can be evaluated on the basis of how well they perform, objectively and subjectively. The question of whether to use rules, or sums-of-products models [17], or neural networks [2] or CART models [11] can be left to empirical investigations. There is no "truth" other than the basic principle that models with greater structure make greater assumptions, and require more data to train. On the other hand more structured models are better at generalising to new situations. These are all just standard issues in machine learning where the choice of model depends on what can be supported by the data. With the most data, we end up with a rule, with the least we must fall back on case-by-case reasoning.

Perceptual unit selection: the objective of simulation is to fool a human listener, so it is imperative that we quantify the abilities of a listener to judge mismatches in spectrum, timing or phonetic quality. This perceptual information can then be incorporated into the unit selection process. Eventually we will be able to define synthesis as a mathematical optimisation problem in which the generated signal is the one most likely to be considered natural by a listener.

Extrapolation and interpolation: our corpora will always be too small: they can never encompass all speakers, styles, contexts, prosodies or emotions. Signal processing techniques will always be required to extrapolate the data we have to a new situation, or to generate new variants by interpolating between known instances. This manipulation of signals need not be confined to fundamental frequency and duration. Statistical signal processing techniques could be applied to change speaker, voice quality, accent or even speaking style. These are just mapping problems that can always be addressed to some extent from the analysis of suitably labelled corpora.

4. LIMITS TO SPEECH SIMULATION, THE ROLE OF SPEECH SCIENCE

How far can we go with speech simulation without doing speech science? Clearly we have already gone a long way: we have systems that are intelligible although they don't have a vocal tract, that can communicate information although they don't understand it themselves, that appear to express a style of speech even though they use no pragmatics. What reasons are left for studying humans?

Signal processing schemes for interpolation and extrapolation of the speaker, prosody or style of recordings will have limits. It is not necessarily true that a recording of one speaker can be convincingly mapped to another, or that one style can be interpolated from two others. Although we know that there are different speakers and styles in the world, we have no evidence that a recording of one speaker in one style can always be mapped to another speaker or style.

Listeners are sensitive to the intentions of a speaker, perhaps through some coherence in how information is encoded in the signal. We can gauge quite easily whether we are listening to a recorded announcement rather than a person talking solely to us. It is possible that speech simulation systems that do not understand what they are saying, and have no desire to communicate, will always sound like announcement systems.

Building optimal unit selection functions requires knowledge of how a human listener will process the simulated speech signal. Oddly, this will require a lot more knowledge of human speech perception and language processing than we know now. It might be argued that a science of speech communication will only arise when we start to take seriously the goal to produce simulations that satisfy a listener. For the first time we will be taking into account in the generation of a message how the message will be received.

When a human speaker produces an utterance there are many interactions between content, style, emotion, physiological state and the processes of articulation. It is not necessarily the case that these influences can be disentangled and modelled. Iinformation about the underlying causes of changes in prosody, articulatory precision or voice quality may be lost or distorted in speech production. There may not be enough information in any practically-sized corpus to systematise the relations.

These possible weaknesses of speech simulation can themselves be targets for scientific investigation. Thus I see the scientific input to speech synthesis going in two directions: firstly in support of better simulations; secondly in construction of a computational model of a human talker.

The first kind of research looks at the way in which simulations fail – where does simulated articulation, voice quality or prosody break down? What additional information needs to be supplied by the message generation component to select appropriate prosody, and how should the signals in the corpus be labelled? What are the domains of coarticulation and how much coarticulation is perceptible? What spectral discontinuities are most noticeable, and how good are listeners at judging rhythm? These are interesting questions in simulation science but are not about how humans talk.

The second kind of research looks at building a system with a vocal tract, with hearing, with proprioceptive feedback, with the ability to mimic what it hears and monitor its own performance. A system with an innate ability to learn, to want to learn and to want to communicate. In other words a computational talker which can be a testbed of how a human speaks and how a human learns to acquire speech. In contrast to Shadle and Damper [12], I am not convinced that we need to program in the articulations from many imaging studies. It seems imperative to me that the system learns how to articulate from experience with making sounds, and at the same time learns the phonological organisation underlying those sounds.

5. CONCLUSION

I have tried to argue in this paper that corpus methods have changed the nature of research in speech synthesis. Research into unit selection, prosody manipulation and statistical processing of text have little in common with how humans produce speech. This causes tension within the field of speech science. I would like to argue that speech simulation is a proper area of research in its own right, with well defined goals. It is reasonable that part of speech science should be concerned with building better simulations whether or not this contributes to understanding of human production. But if we are to take speech science seriously, then we must also look into building computational models of human talkers, ones that faithfully model the actual processes of speech as far as our theories are able to describe. It is unlikely that such articulatory synthesis will make any contribution to speech simulation in the medium term; but it is a proper goal for science.

ACKNOWLEDGEMENTS

The author wishes to thank the members of the COST-258 project for encouraging the development of the ideas presented in this paper.

REFERENCES

[1] Breen, A. P. and Jackson, P., “A phonologically motivated method of selecting non-uniform units”, in International Conference on Speech and Language Processing, 1998

[2] Campbell, W.N., “Syllable-based segmental duration”, in G. Bailly, C. Benoit, and T. Sawallis, editors, Talking machines: Theories, models, and designs, pages 211-224. Elsevier, 1992.

[3] Damper, R.I., Stanbridge, C.Z., Marchand, Y., “A Pronunciation-by-Analogy Module for the Festival Text-to-Speech Synthesiser”, in 4th ISCA Workshop on Speech Synthesis, 2001.

[4] Fackrell, J.W.A., Vereecken, H. , Martens, J.-P., Van Coile, B. "Multilingual prosody modelling using cascades of regression trees and neural networks", Proceedings of EuroSpeech-99, 1835-1838, 1999.

[5] Flanagan, J.L., Speech Analysis Synthesis and Perception. Springer-Verlag, New York, 1965.

[6] House, J., Dankovicova, J., and Huckvale, M., "Intonation modelling in ProSynth: an integrated prosodic approach to speech synthesis", Int. Congr. Phonetic Sciences, 1999.

[7] Llitjos, A., and Black, A., “Knowledge of language origin improves pronunciation of proper names”, Proceedings of EuroSpeech-01, 1919-1922, 2001.

[8] Hitzeman, J., Black, A.W., Taylor, P., Mellish, C., Oberlander, J., “On the Use of Automatically Generated Discourse-Level Information in a Concept-to-Speech Synthesis System”, International Conference on Speech and Language Processing, 1998.

[9] Moulines, E., and Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, 9, 453-467, 1990.

[10] Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovicova, J., Heid, S., “ProSynth: an integrated prosodic approach to device-independent natural-sounding speech synthesis”, Computer Speech and Language, 14, 177-210, 2000.

[11] Riley, M., “Tree-based modelling of segmental durations”, in G.Bailly, C.Benoit (eds) Talking Machines. Theories, Models and Designs, 265-273, North Holland, 1992.

[12] Shadle, C.H., Damper, R.I., “Prospects for Articulatory Synthesis: A Position Paper”, in 4th ISCA Workshop on Speech Synthesis, 2001.

[13] Stevens, K.N., Bickley, C.A., and Williams, D.R., "Control of a Klatt synthesizer by articulatory parameters." Proc. International Conference on Spoken Language Processing, 183-186, 1994.

[14] Stork, D. (Ed), Hal’s Legacy, 2001’s computer as dream and reality, MIT Press, 1997.

[15] Stylianou, Y., Cappé, O., and Moulines, E., “Continuous probabilistic transform for voice conversion”, IEEE Trans. Speech and Audio Processing, 6(2):131-142, March 1998.

[16] Taylor, P.A., and Black, A.W., “Concept-to-speech synthesis by phonological structure matching”, Proceedings of EuroSpeech-99, 623-626, 1999.

[17] Van Santen, J.P.H., “Assignment of segmental duration in text-to-speech synthesis”, Computer Speech and Language, 8, 95-128, 1994.