A READING LIST OF RECENT ADVANCES IN SPEECH SYNTHESIS

Simon King

The Centre for Speech Technology Research, University of Edinburgh, UK
Simon.King@ed.ac.uk

ABSTRACT

This is a companion paper to my keynote talk at ICPhS 2015. It provides a guide to help readers familiarise themselves with recent advances in speech synthesis, with an emphasis on approaches that might provide useful tools to investigate speech, particularly by constructing experimental stimuli for perceptual experiments.

Keywords: Text-to-speech; speech synthesis; speech manipulation

1. INTRODUCTION

Manipulated speech is an essential tool for experimental work in the speech and hearing sciences, helping us to answer a host of questions about both production and perception. Whilst our understanding of speech marches inexorably onwards, the tools available have not kept up. The laboratory phonologist is still faced with difficult decisions every time a new experiment is designed, and often has to make compromises in the materials to be used simply because of the effort required to construct them.

In this paper, I will survey some techniques that are currently available from the field of text-to-speech synthesis. Because of what has become possible quite recently in terms of reasonably natural and highly intelligible synthetic speech, it seems a good time to ask which of the technologies behind synthetic speech generated from text could be repurposed and applied to the scientific investigation of speech itself.

Section 2 offers some pointers to the main methods currently in use. From this, we can identify component technologies that obviously have some potential as tools for scientific investigations: waveform concatenation; vocoders; statistical parametric models.

The remainder of this paper takes the form of a reading guide to recent literature on speech synthesis, always with the question in mind "Could this be useful for speech science?" Some potential advantages of uses these technologies, compared to traditional manipulation methods, are: control over individual acoustic aspects of speech; ability to produce stimuli that human talkers cannot; many different voices; much larger variety (and sheer quantity) of stimuli. The suggested readings have been selected as good entry points to the literature and the number of citations is kept under control to focus the reader on just one or two useful readings per topic. To discover additional readings, search for recent papers that cite the items mentioned here.

2. BACKGROUND

Text-to-speech (TTS) research, driven by the demands of mainstream commercial applications, has delivered a sequence of remarkable improvements in naturalness and intelligibility. These advances can mostly be attributed to the widespread adoption of statistical modelling, which has now largely replaced the somewhat hand-crafted approach of unit selection, at least on the research agenda.

2.1. Unit selection

Traditional unit selection speech synthesis, in which novel utterances are created by re-arranging fragments of pre-recorded speech according to carefully crafted linguistic and acoustic cost functions, is still in widespread commercial use. This technology can – when well engineered and expertly tuned – produce some of the most natural synthetic speech available even today. Commercial legacy unit selection systems will no doubt remain operational for many years.

The problem with traditional unit selection is not the idea of concatenating waveform fragments, but rather the fragility of the linguistic ('target') and acoustic ('join') cost functions, which are only able to reliably select natural-sounding sequences of units if the speech database is accurately labelled and is consistent in terms of recording quality and speaking style. The most critical components of a unit selection system are therefore the database, and the target cost function that determines how suitable a unit from the database is for the different linguistic context of the utterance being synthesised.

What to read:

- The best available textbook on speech synthesis is by Taylor [15], which offers a comprehensive treatment.

2.2. The statistical parametric approach

In so-called HMM-based (Hidden Markov Model-based) speech synthesis, no waveforms are stored. Instead, the speech database is used to train a set of context-dependent phone models, which are used to drive a vocoder at synthesis time.

Although the models are indeed HMMs, it is actually better to think of this method of synthesis as a large regression tree which queries the linguistic context ("Is this a vowel?"; "Is there a nasal to the left?"; "Are we in a stressed syllable?"; etc) of the current phone being generated, and arrives at a statistical description (i.e., mean and covariance) of the vocoder parameters. Each 'stream' of vocoder parameters (one stream for the spectral envelope, another for F0, etc) is predicted by a separate regression tree, which means that the most appropriate questions can be selected in each case.

Once we understand that the hard work is being done by a regression tree, it is only a small step to replace the tree with any other general-purpose regression model, such as a neural network.

What to read:

- A non-technical introduction to statistical parametric speech synthesis can be found in [5].
- A deeper and more technical introduction is available in [19].
- The neural network approach is relatively new, but has already exploded in popularity, not least because it is so closely related to the HMM-based approach. It is too early for a comprehensive review paper at this time, but [7] gives a snapshot of the field.

2.3. Hybrid methods

Evidence that waveform concatenation itself is still an attractive proposition comes from so-called hybrid systems, in which a statistical parametric model guides the selection of units. In effect, a complete statistical parametric (a.k.a. HMM-based) system is first built. But instead of using it to drive a vocoder, its regression tree is used to replace the target cost function. The tree predicts acoustic features from linguistic context, and then waveform fragments with similar acoustic properties are retrieved from the database. The reasons that this outperforms a hand-tuned function are that it is not only learned from data, but it is also specific to the current speaker and the particular recorded speech database being used.

What to read:

- Microsoft's term for their hybrid approach is 'trajectory tiling' [10] which neatly captures the idea of 'sketching' a kind of 'wireframe' of speech parameters, and then 'tiling over' it with waveform fragments, by analogy with photo-realistic rendering in computer graphics.

2.4. Naturalness and intelligibility

Intriguingly, statistical parametric systems produce the most intelligible speech, but are not rated as sounding particularly natural by listeners. Conversely, unit selection systems, which listeners agree sound most natural, are usually significantly less intelligible than their statistical parametric cousins. Hybrid systems can match or exceed the naturalness of unit selection, whilst coming closer to statistical parametric in terms of intelligibility.

This surely tells us something interesting about speech perception ... but what? Anyone caring to investigate that can find a ready-made set of materials, complete with listener ratings of naturalness and their typed-in responses to the intelligibility test, as part of the distributed output of the Blizzard Challenge.

What to read:

- The Blizzard Challenge Website links to all the papers from the Challenge, including annual summary papers, as well as to the materials mentioned above: http://www.synsig.org/index.php/Blizzard_Challenge

3. VOCODERS

Statistical parametric synthesis relies on the use of a vocoder to convert waveforms into a parametric form suitable for modelling, then to convert model-generated speech parameters back in to waveforms during synthesis. A wide variety of vocoders exist but here it is only necessary to consider STRAIGHT, because it is the most widely used, and also to mention a quite different class: sinusoidal vocoders.

3.1. STRAIGHT

This vocoder has been used extensively in statistical parametric speech synthesis. The goal of STRAIGHT is to achieve source-filter separation. However, STRAIGHT is *not* strictly a source-filter model. Rather, it models the *spectral envelope* and has no explicit vocal tract filter model. The principal strength of STRAIGHT is in the analysis phase,

where speech waveforms are converted to vocoder parameters. Instead of adopting a particular model of the vocal tract (e.g., an all-pole filter with a fixed number of resonances), STRAIGHT simply makes the assumption that the spectral envelope is smooth in both frequency and time, and uses a clever pitch-adaptive window to reduce interference from the harmonic structure when estimating that smooth envelope. To resynthesise the speech, a filter must be constructed from the spectral envelope; this is excited with a source signal that mixes a phase-manipulated pulse train with shaped noise. A software implementation of STRAIGHT is available, along with a graphical user interface for producing a continuum of stimuli between two natural samples, via morphing of the vocoder parameters.

How it might be used:

- High-quality time and pitch modifications.
- Spectral envelope transformations (e.g., formant frequencies).
- Morphing between two natural samples, to create a continuum.

What to read:

- There are relatively few papers on STRAIGHT, but [4] is the place to start, since it outlines the philosophy behind this vocoder.

3.2. Sinusoidal / harmonic-plus-noise vocoders

Whilst the majority of speech signal models – including STRAIGHT – used in speech synthesis have a source-filter architecture or equivalent, another class of vocoders is available. These vocoders take a pragmatic view and attempt to model the signal directly, without explicit reference to any model of speech production. The signal is assumed to be the sum of a deterministic part (the harmonic structure, modelled as a set of sinusoids) and a stochastic part (the noise components). Developments on this basic idea have led to vocoders that are more transparent than STRAIGHT and enable very high quality modifications with fewer artefacts than STRAIGHT. There is a significant downside: the number of parameters needed to represent the signal is both large, and variable, making these models problematic for direct use in statistical parametric text-to-speech.

Whilst viewing the speech signal as a sum of sinusoids and some noise is not perhaps the most natural, from either a speech production or a speech perception point of view, the excellent quality that this class of vocoders can achieve makes them worthy of consideration for use in the laboratory, wherever natural speech must be manipulated.

How it might be used:

- Very high-quality time and pitch modifications.
- Future potential to do anything that STRAIGHT can do, with higher quality.

What to read:

- [6] is the paper that first introduced the idea of summing harmonic and noise components to model speech signals.
- Further developments from that starting point include [14, 9, 1].
- [3] compares many vocoders in a perceptual test.

4. PARAMETRICALLY-CONTROLLABLE SPEECH

4.1. Model adaptation

The ability to modify the parameters of statistical parametric models in principled ways is a significant factor in their widespread use. Transformations to global fundamental frequency and speaking rate are trivial, simply by modifying the values of the corresponding model parameters (e.g., moving the mean value up or down by some fraction of a standard deviation). But far more sophisticated transformations are possible, by applying a different transform to particular subsets of model parameters. Crucially, the transforms are learned from example *adaptation data*, and so previously-trained models can be made to 'imitate' that small sample of adaptation data.

A logical extension of STRAIGHT's ability to interpolate between two natural samples (e.g., the endpoints of a desired continuum) is to interpolate between complete statistical models, or sets of adaptation transforms. Indeed, extrapolation is also possible. In effect, a continuum of complete text-to-speech systems can be created, perhaps along a dimension such as emotion, speaking style, or speaker identity. In this way, any sentence can be generated in any style, or with any speaker identity.

How it might be used:

- Modifying global voice characteristics.
- Creating stimuli to match speaker identity of existing recordings, without needing the original speaker.
- Creating extrapolated speaking styles, beyond natural limits.

What to read:

- [17] gives a technical description of an advanced adaptation technique and contains plenty of experimental results for adapting to new speakers.
- [16] gives an example of how adaptation can be used to create a voice from relatively short recordings (of a child, in this case), made in imperfect conditions.

4.2. Multiple-Regression Hidden Markov Model

Model adaptation works very well, but the control over the output is implicit, via the adaptation data. There are no explicit controls. The Multiple-Regression Hidden Markov Model (MR-HMM) provides for external control parameters. These can, in principle, be any external variable that is known for the training data. The most convincing demonstration of this approach uses articulator positions, measured using Electromagnetic Articulography, as the control 'knobs'. It is possible to affect the synthetic speech in terms of these controls, for example moving the tongue at synthesis time to create sounds that may not even have been represented in the training data.

How it might be used:

- Explicit control over any aspect of speech that can be captured in parallel with the speech data, or labelled afterwards. Obvious examples would be articulator positions or formants.

What to read:

- The key paper demonstrating control via articulation is [8].

4.3. Cluster-adaptive Training

One problem with many of the methods for adaptation and control already mentioned is that they are in some sense too powerful. That is, they are able to generate sounds outside the bounds of normal speech. In some limited circumstances, this might be desirable (e.g., modest amounts of extrapolation), but in general it can lead to unstable systems that create audible artefacts and unnatural speech. Cluster Adaptive Training (CAT) solves that problem by limiting the space of possible adapted models to be a linear interpolation of a set of clustered models. The space of models thus possible is determined by the clustered models, and so the system can be configured to give control over speaker identity, or speaking style, or anything else represented in the training data. CAT can also provide external controls, somewhat like the MR-HMM, but these are not explicitly labelled but are rather discovered from the data during training.

How it might be used:

- Control over any aspect of speech represented in a diverse training corpus, such as speaker, speaking style, etc.
- Simultaneous control over multiple aspects via low-dimensional intuitive user controls (e.g., exposed to the user as a 2-dimensional joystick), albeit without labels.

What to read:

- The original use of CAT [2] was for automatic speech recognition, and one example of its application to speech synthesis can be found in [18].

5. AUTOMATED WAVEFORM EDITING

Whilst parametric methods offer a great deal of control over the speech signal, they are limited by the transparency of the vocoder. That is, even under conditions of little or no modification, the naturalness is still impaired. Waveform concatenation is still the only technique that can avoid this. The 'no modification' condition is simply be playback of complete recorded utterances.

We know that it is hard to hand-tune to cost functions in a unit selection system, and that they are are quite sensitive to errors in database labelling. In other words, the system fails to automatically choose the most natural-sounding sequence of waveform fragments, because of those limitations. In the heyday of commercial unit selection development, it was quickly realised that the quality of synthetic speech obtained from a fixed pre-recorded database could quite easily be improved through a little manual intervention during synthesis. In other words, instead of using the automatically-chosen unit sequence, another sequence would be manually selected, perhaps taking the second- or third-best choices for some units. Of course, this wouldn't work for a live text-to-speech application, but has applications in preparing recorded announcements.

How it might be used:

- Semi-automatic construction of spliced speech from a (potentially very large) corpus.

What to read:

- Interactive synthesis [11] was developed commercially into a tool for 'prompt sculpting' aimed at preparing prompts for telephone dialogue systems [12] and was eventually patented [13].

6. REFERENCES

[1] Degottex, G., Stylianou, Y. 2013. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Transactions on Audio, Speech and Language Processing* 21(10), 2085–2095.

[2] Gales, M. 2000. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Audio, Speech and Language Processing* 8(4), 417–428.

[3] Hu, Q., Richmond, K., Yamagishi, J., Latorre, J. 2013. An experimental comparison of multiple vocoder types. *8th ISCA Workshop on Speech Synthesis* Barcelona, Spain. 155–160.

[4] Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27(3-4), 187–207.

[5] King, S. 2011. An introduction to statistical parametric speech synthesis. *Sadhana* 36(5), 837–852.

[6] LaRoche, J., Stylianou, Y., Moulines, E. 1993. Hnm: a simple, efficient harmonic+noise model for speech. *Applications of Signal Processing to Audio and Acoustics, 1993. Final Program and Paper Summaries., 1993 IEEE Workshop on* 169–172.

[7] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H., Deng, L. 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine* 32(3), 35–52.

[8] Ling, Z.-H., Richmond, K., Yamagishi, J. 2013. Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression. *IEEE Transactions on Audio, Speech and Language Processing* 21(1), 207–219.

[9] Pantazis, Y., Rosec, O., Stylianou, Y. 2011. Adaptive AM–FM signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech and Language Processing* 19(2), 290–300.

[10] Qian, Y., Soong, F., Yan, Z.-J. 2013. A unified trajectory tiling approach to high quality speech rendering. *IEEE Transactions on Audio, Speech and Language Processing* 21(2), 280–290.

[11] Rutten, P., Fackrell, J. 2003. The application of interactive speech unit selection in TTS systems. *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*.

[12] Rutten, P., Talkin, D. 2004. rVoice studio and ActivePrompts. *5th ISCA Workshop on Speech Synthesis* Pittsburgh, PA, USA.

[13] Rutten, P., Taylor, P. Sept. 3 2013. Method and apparatus for sculpting synthesized speech. US Patent 8,527,281.

[14] Stylianou, Y. 1996. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis Ecole Nationale Supérieure des Télécommunications.

[15] Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

[16] Watts, O., Yamagishi, J., King, S., Berkling, K. 2010. Synthesis of child speech with HMM adaptation and voice conversion. *IEEE Transactions on Audio, Speech and Language Processing* 18(5), 1005–1016.

[17] Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S. 2009. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing* 17(6), 1208–1230.

[18] Zen, H., Braunschweiler, N., Buchholz, S., Gales, M., Knill, K., Krstulovic, S., Latorre, J. 2012. Statistical parametric speech synthesis based on speaker and language factorization. *IEEE Transactions on Audio, Speech and Language Processing* 20(6), 1713–1724.

[19] Zen, H., Tokuda, K., Black, A. W. 2009. Statistical parametric speech synthesis. *Speech Communication* 51(11), 1039 – 1064.