This is a draft chapter for the book "Natural Speech Technology: Statistical and Machine Learning Approaches to Speech Recognition and Synthesis".

The chapter text was mainly written prior to and during November 2016, and reflects the shape and general understanding of statistical speech synthesis at that time. The draft was edited in July 2017 for length and edited again in July 2018 to function as a stand-alone document, as presented here. This version was compiled on July 31, 2018.

1 Analysing Shortcomings of Statistical Parametric Speech Synthesis

Gustav Eje Henter, Simon King, Thomas Merritt[‡] and Gilles Degottex

1.1 Introduction

Even the best statistical parametric speech synthesis (SPSS) produces output that is noticeably worse than natural speech, whether measured in terms of quality, naturalness, speaker similarity, or intelligibility in noise. These *shortcomings* have been attributed to a multitude of causes, and the literature is awash with solutions to various supposed problems with the statistical parametric approach. Yet, somewhat surprisingly, the hypothesised problem is often not clearly defined, or no empirical evidence us provided to confirm – or quantify – its contribution to imperfections in the synthetic speech. Across the literature as a whole, there is surprisingly little work exploring which of the many potential problems are perceptually the most important; this would, of course, be useful knowledge.

The conventional arguments in favour of the statistical parametric approach are, in contrast, well rehearsed and *are* supported by plenty of experimental evidence: good intelligibility, robustness to imperfect data, ability to adapt the model, and so on.

If we take, as is widely accepted in the field, *naturalness* as our principal measure, then evidence that synthetic speech is inferior to natural speech recordings is overwhelming. The simplest and clearest evidence comes from the the long-running Blizzard Challenge, summarised in (King 2014). Year after year, we see that no synthesiser is ever judged to be as natural as recorded speech.

In contrast to naturalness, the Blizzard Challenge shows that impressive progress has been made in *intelligibility*, where – in quiet conditions at least – some synthesisers are as good as recorded speech. This is not (yet) generally the case in non-quiet listening situations. The Hurricane Challenge (Cooke, Mayo & Valentini-Botinhao 2013) provides convincing evidence that synthetic speech intelligibility in the presence of additive noise remains substantially inferior to that of recorded speech.

The Blizzard and Hurricane Challenges both compare different synthesis approaches on the same data, demonstrating that it is the specific implementation and assumptions of each synthesiser that are responsible for differences in naturalness or intelligibility. There are apparently only two or three different waveform generation technologies represented in the entries the these challenges: statistical parametric systems employing a vocoder, unit selection employing waveform concatenation, and a *hybrid* variant of unit selection that uses an internal statistical parametric model. There are clear trends,

[‡] Work done prior to joining Amazon.

consistently across many years of the challenge: 1. statistical parametric systems are generally the most intelligible; 2. unit selection systems are generally the most natural-sounding; 3. hybrid systems can achieve the naturalness of unit selection whilst approaching the best intelligibility. Strictly speaking, we can only say say that the choice of waveform generation technology *correlates* with intelligibility and naturalness; we can't make the stronger claim of *causality*.

Because individual system details are seldom open knowledge, and because entire systems are evaluated "end to end", it is not straightforward to attribute successes and shortcomings to specific elements in each approach. This is particularly true about the front-end text processor, partly because much recent research has neglected the effect of this component and focussed much more on acoustic modelling and waveform-generation technology.

In this chapter, we analyse some of the shortcomings of SPSS. Since there is at least a correlation between waveform generation technology and synthetic speech naturalness, we start with vocoding. This is followed by a description – with an example application – of a general methodology for quantifying the effect of any of the many assumptions, design choices, and (possibly inherent) limitations of SPSS. The example application includes measuring the shortcomings of vocoding, relative to other limiting factors such as the statistical model.

1.2 Vocoding

The role of the vocoder in SPSS is to provide a representation of the speech signal that is suitable for statistical modelling and at the same time from which a waveform can be generated. Therefore, vocoder design inevitably involves a trade-off between the two. For example, dimensionality reduction and an approximately decorrelating transform, e.g., retaining only low-order mel-cepstrum coefficients, are widely applied, either within the vocoder, or to its parametrisation.

Vocoded (analysis-synthesis) speech is the assumed upper-bound of the quality achievable from a SPSS system in the limit of a highly accurate statistical model, and it is therefore of interest to measure how much the vocoder alone limits the achievable quality of all systems that employ one.

The very act of parametrising a speech signal, which is itself a non-linear combination of interacting sound sources and sound-shaping processes, creates many challenges. For example, it is known that the glottal source has a particular amplitude spectrum which nevertheless most vocoders assume is flat, combining all spectral envelope modelling into a single component that also handles the amplitude spectrum of the Vocal Tract Filter (VTF). However, the true glottal source amplitude spectrum varies with F0. Most current vocoders take a simple approach, assuming the VTF is independent of F0, and therefore with no separate modelling of glottal source amplitude spectrum, non-periodic sound generated in the glottis, or the glottal source phase spectrum.

Another example is the binary voicing decision (speech is either voiced or unvoiced) that many vocoders incorporate, which is arguably an oversimplification. In natural speech, at transitions from unvoiced to voiced, or vice versa, it is frequently the case that the so-called deterministic (i.e., voiced, periodic) speech component somewhat gradually commences or fades away. As a consequence, the way in which voicing is handled by a vocoder may lead to differences in synthetic speech quality (Latorre, Gales, Buchholz, Knill, Tamurd, Ohtani & Akamine 2011, Yu & Young 2011, Degottex & Erro 2014).

As a final example of the many issues in vocoding, the seemingly random timing of glottal pulses in creaky voice, also called vocal fry (Laver 2009), is generally poorly captured in current vocoders because they use a perfectly periodic pulse train excitation signal for all voiced sounds. Irregularities in the mechanical vibration of the vocal folds are commonplace (Drugman, Kane & Gobl 2014). Current speech analysis techniques are prone to confusing these irregularities with simple additive noise, which leads to vocoders producing speech with a perceived hoarse voice quality instead of a creaky quality. Similarly, breathy voice (Laver 2009, Ishi, Ishiguro & Hagita 2010) can only by synthesized by the simple addition of noise, whereas in actual speech production the shape of the glottal pulse might be also varying rapidly, producing a signal that cannot be approximated by a pulse train plus noise.

Even though there are some advanced techniques for creak detection (Kane, Drugman & Gobl 2013, Drugman et al. 2014) and measures of pulse variation (Degottex & Erro 2014) it is not obvious how to build a vocoder that takes advantage of such features in a statistical modelling framework, without substantially increasing the dimensionality of the representation.

The standard speech parametrisation setup used in statistical parametric speech synthesis is outlined in (Zen, Tokuda & Black 2009). However, for the reasons above, and others, this configuration has a degrading effect on perceived quality of speech (Merritt, Raitio & King 2014, Henter, Merritt, Shannon, Mayo & King 2014, Merritt, Latorre & King 2015). Furthermore, the amount of degradation may differ markedly between speakers (voices), and one vocoder or another may perform better or worse for any particular speakers; cf. (Babacan, Drugman, Raitio, Erro & Dutoit 2014) for singing. These quality variations suggest a notion of “vocodability”, the consequence of which is a (perhaps undesirable) bias when selecting a speaker for a TTS corpus, towards a voice that suffers minimal degradation at the hands of the vocoder.

Approaches to improving the modelling of speech signals for SPSS can be described in three distinct categories: source-filter parametrisations, sinusoidal parametrisations and non-parametric approaches. A further comparison of selected source-filter and sinusoidal parametrisation methods can be found in (Hu, Richmond, Yamagishi & Latorre 2013).

1.2.1 Source-filter parametrisation

The idea that source and filter can be separated is a simplification of the voice production mechanism. Coupling exists: when articulator positions change, this not only changes the VTF, but also has an effect on the glottal pulse spectrum (Fant & Lin 1987); the more open the glottis, the wider the formant bandwidths (Hanson 1997); etc. So, source and filter are obviously dependent and correlated, since they are the intertwined consequences of articulator movement. The acoustic consequences of source and of filter are therefore also not entirely separable. We examine the impact of independent modelling of source and filter in section 1.4.6.

The other obvious limitation of the source-filter model is that the location of the sound source is not always the glottis, but can be elsewhere in the vocal tract, such as the constriction for a fricative, or the closure and release of a plosive. For these reasons, we should state clearly that source-filter models are only models of speech signals, *inspired* by speech production mechanisms, but not faithful to them.

One line of research that may eventually lead to an alternative solution, one that *is* faithful to speech production mechanisms, is so-called physical modelling; that approach is still a long way from offering this solution: the models are simplified and incomplete, and fitting their parameters to natural speech signals cannot be done reliably. Therefore, the vast majority of SPSS systems use models of the speech signal, not of the production process.

Staying within the source-filter signal modelling paradigm, it may be possible to improve modelling accuracy with a more sophisticated source. As mentioned earlier, the glottal source magnitude and phase should ideally not be assumed constant and flat, although this is what STRAIGHT (Kawahara 2006) and most other vocoders assume. Natural glottal pulses actually have a non-minimum-phase spectrum which therefore cannot be predicted from the amplitude spectrum. This is also true of the VTF.

Many approaches have been suggested to improve the model of the glottal source for SPSS (Klatt & Klatt 1990, Raitio, Suni, Yamagishi, Pulakka, Nurminen, Vainio & Alku 2011, Cabral, Renals, Richmond & Yamagishi 2007, Degottex, Lanchantin, Roebel & Rodet 2013) either by estimating the glottal pulse waveform (Raitio et al. 2011) or the parameters of an analytical model of it (Cabral et al. 2007, Degottex et al. 2013). In addition to attempting to solve the issues mentioned above, the analytical model approach also provides parameters that are not merely generic signal-based features, but more closely related to the underlying physical system. For example, the *Rd* coefficient (Fant 1995) links the amplitude and phase spectra of the glottal source in a way that is governed by physical constraints.

1.2.2 Sinusoidal parametrisation

The other main type of signal analysis used in SPSS is the sinusoidal model (McAulay & Quatieri 1986, Stylianou 1996), although this is most often used only as an intermediate representation, with a subsequent dimensionality-reducing parametrisation being required for statistical modelling. Sinusoidal models can be seen as sparse representations of the speech signal:

$$s(t) = \sum_{k=-K}^{K} a_k(t) \cdot e^{j\phi_k(t)} \tag{1.1}$$

where K is the number of sinusoidal components, with amplitude $a_k(t)$ and phase $\phi_k(t)$ (Degottex & Stylianou 2013). The spectral amplitude envelope can be constructed from

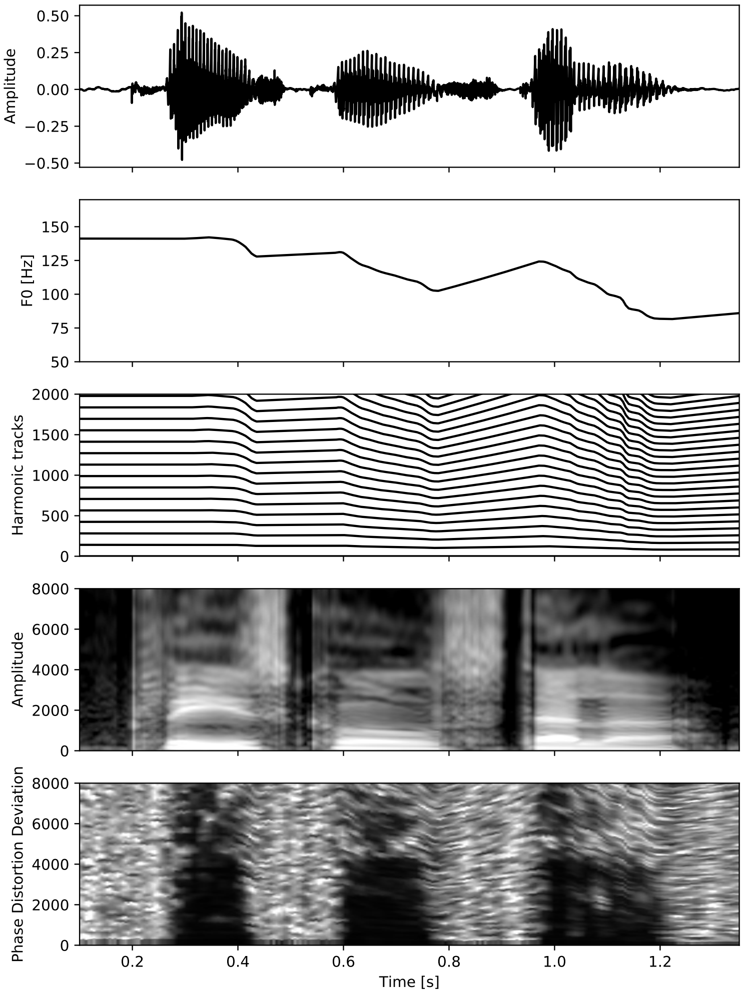


Figure 1.1 An example of parameters estimated by the HMPD vocoder. From top to bottom: waveform; continuous fundamental frequency F0 curve; harmonic tracks used for the estimation of $a_k(t)$ and $\phi_k(t)$; the amplitude spectral envelope modelling $a_k(t)$ (the lighter the colour the louder); the Phase Distortion Deviation (PDD) modelling the random component of $\phi_k(t)$ (the lighter the colour the noisier)

the amplitude parameters (El-Jaroudi & Makhoul 1991) while a phase envelope can be found from the phase parameters (Agiomyrgiannakis & Stylianou 2009, Degottex & Erro 2014).

Because the reconstruction quality of sinusoidal models is very high (Degottex &

Stylianou 2013), attempts have been made to directly model the parameters (Hu, Yamagishi, Richmond, Subramanian & Stylianou 2016) without requiring further parametrisation (e.g., as a spectral envelope represented by a truncated mel cepstrum). Direct modelling of sinusoidal model parameters falls somewhere between the fully parametric source-filter approach above (e.g. STRAIGHT) and the non-parametric approach described next.

An open issue is how to manage the dimensionality of sinusoidal models (Hu et al. 2016), since there are a large number of highly-correlated parameters. A further complication is that the number of sinusoids below the Nyquist frequency varies over time.

1.2.3 Non-parametric representation

Given that the vocoder limits naturalness, there is interest in integrating some or all of the vocoder signal processing into the statistical model, essentially enabling the parametric speech representation to be learned and improved, rather than being static (even if carefully engineered). Automatic Speech Recognition and other signal processing applications have undergone significant changes quite recently to using lower-level features. For speech synthesis, one example is predicting the full-detail, high-dimensional STRAIGHT spectrum rather than working with mel-generalised cepstral coefficients (MGCs), e.g., (Ling, Deng & Yu 2013). This removes a potentially restrictive dimensionality reduction, at the expense of having to learn to model highly correlated features.

Going further, there are ways to directly generate a waveform without reconstructing it from an engineered intermediate representation such as the spectral envelope. One step towards direct waveform generation was the generation of glottal pulse waveforms glottal source signal in speech (Raitio, Lu, Kane, Suni, Vainio, King & Alku 2014), even down to the level of individual samples (Juvela, Bollepalli, Airaksinen & Alku 2016), although these were then used in a source-filter model. Another way to generate a waveform is to use the output from an SPSS system to control the selection of units in a unit selection system: hybrid synthesis, described briefly below. Attempts have also been made to statistically model the spectral properties directly from waveforms instead of passing through an explicit estimate of the spectral envelope (Tokuda & Zen 2015, Tokuda & Zen 2016).

Most recently, WaveNet (van den Oord, Dieleman, Zen, Simonyan, Vinyals, Graves, Kalchbrenner, Senior & Kavukcuoglu 2016) suggests that direct waveform synthesis is not merely a theoretically interesting concept, but is capable of very high output quality. This performance currently comes at prohibitive computational cost, as well as concerns over data quantity requirements and patent status, which may limits the usefulness of this particular approach.

An alternative technique to obtain high segmental quality is to use waveform segments from the training database for signal generation. Hybrid synthesis (unit selection driven by SPSS) is well-established and in widespread commercial use. Hybrid systems have made a strong showing in the Blizzard Challenge ever since their arrival (Ling, Qin, Lu, Gao, Dai, Wang, Jiang, Zhao, Yang, Chen & Hu 2007, Ling, Lu, Hu, Dai & Wang 2008, Lu, Ling, Lei, Wang, Zhao, Chen, Hu, Dai & Wang 2009, Jiang, Ling, Lei, Wang, Heng, Hu, Dai & Wang 2010), the approach has subsequently grown in popularity there (King 2014), and Open Source implementations are becoming available (Merritt, Clark, Wu, Yamagishi & King 2016). Other permutations of SPSS and unit selection include multiform synthesis (Pollet & Breen 2008), where a sequence of SPSS-generated speech and recorded waveform units are concatenated, or manipulating recorded speech prototypes to match predictions from SPSS (Espic, Valentini-Botinhao, Wu & King 2016).

1.2.4 Summary

The near future of speech synthesis will certainly involve continued attempts to move beyond the use of vocoders. As deep learning improves the state of the art in many areas of spoken language processing, including TTS, it becomes more obvious how to replace traditional signal processing with a learned pipeline. By recasting some or all of the acoustic feature processing as layers of neural network (for example), end-to-end optimisation of acoustic model and signal representation becomes apparently straightforward. For example, (Takaki, Kim, Yamagishi & Kim 2015) investigated the many roles that a DNN might fulfil in a TTS system, and WaveNet (van den Oord et al. 2016) demonstrated one way to directly synthesise a waveform, although optimising the DNN loss function at the waveform sample level would seem to be only very loosely related to perceived error.

Burying all of the speech signal processing inside a statistical model might not be desirable for some applications. Traditional representations used in vocoders, such as source and filter, are intuitive and amenable to manipulation. Pitch, for example, can be easily tuned according to listener preference by applying simple scaling to F0. The spectral envelope can be frequency-warped in order modify speaker identity. Duration can be scaled to manipulate speaking rate. Such techniques provide very simple and efficient ways to generate a variety of speakers and styles from a single statistical model.

1.3 Attributing degradations to modelling assumptions by performing selective comparisons

The previous section discussed why vocoded speech (analysis-synthesis) is worse than natural speech. But degradations in synthetic speech – that is, the ways in which synthetic speech is worse than natural speech – are not limited to the vocoder. In general, speech generated from text using a statistical parametric model is more degraded than vocoded speech in terms of signal quality, expressivity, similarity to the original speaker, and intelligibility. Model-generated prosody can be inappropriate or unconvincing. Synthetic speech is generally judged as significantly less natural than vocoded speech, and can be unpleasant to listen to over longer periods (Wester, Watts & Henter 2016).

The overall quality is a consequence of myriad interacting factors. The remainder of section 1.3 describes a general methodology that can be used to tease apart the effects of different modelling assumptions. In section 1.4, we review selected findings from the literature in the light of the described methodology, and ask what they tell us about the effects of common modelling paradigms and assumptions.

1.3.1 Basic comparison methodology

The basic principle for measuring the effects of different speech-synthesis design choices is to contrast the output from two comparable speech synthesis systems; this type of evaluation is widespread. If multiple aspects of a TTS system are changed incrementally in sequence, a chain of different synthesisers is obtained, and the relative severity of the different assumptions and simplifications involved can be studied.

In practice, all evaluations are influenced by the context in which they take place:

1. The **training data** used, discussed in section 1.3.2
2. The **evaluation methodology**, discussed in section 1.3.3
3. The **surrounding model**, discussed in section 1.3.4
4. The **output generation method**, discussed in section 1.3.5

Effects of mathematical modelling assumptions, making up the bulk of the TTS design aspects to be discussed, are covered in section 1.4.

1.3.2 The data

Any given model or approach does not necessarily work equally well for all datasets. Apart from well-known (but not well-understood) variations caused by speaker characteristics, the amount of data obviously has an effect. For a small dataset, a simple decision tree may outperform a more complex neural network, but the complex model may ultimately give best performance once there is enough data. In practice, most investigations involve just a single speaker and a single fixed-size dataset, which limits the generality of their conclusions.

The quality of the data will also have an affect: transcription errors, signal issues including recording noise, reverberation, compression or transmission artefacts, or even problems with the speech articulation (e.g., disordered and dysarthric speech). This is a broad topic which has had limited systematic exploration (Yamagishi, Ling & King 2008, Karhila, Remes & Kurimo 2014, Bollepalli, Raitio & Alku 2013, Creer, Cunningham, Green & Yamagishi 2013).

In general, current commercial practice prevails even in much academic research: fairly large quantities of specially-designed, cleanly recorded, and well-transcribed material is used, even when this leads to a bland and sometimes unnatural speaking style.

1.3.3 The evaluation

Objective comparisons rely on analytic criteria that can be computed and perhaps even optimised automatically, while subjective comparisons generally take the form of listening tests that require careful setup and a group of human evaluators comparing or rating speech stimuli.

Listening tests can produce categorical or numerical outcomes. The former is most common for preference tests or difference detection (discrimination) tests, which tend to be the most sensitive in detecting differences. Tests that directly produce numerical performance measures include mean opinion scores (MOS) (ITU-T 1996) and MUSHRA (ITU-R 2015), of which the latter has been found to be more sensitive (Ribeiro, Yamagishi & Clark 2015). These paradigms are likely be a better choice for assessing the relative severity of different choices in a spectrum of models by measuring effect sizes, rather than only identifying statistically significant differences. It is possible to infer similar information from categorical judgements, such the ratios of similar vs. different judgements between pairs of tested systems. These can be interpreted as relative system similarity, and thus give a picture of listeners' perceptual space. By applying *multidimensional scaling* (MDS) (Borg & Groenen 2005) to similar-different judgements, systems under test can be located in a continuous-valued perceptual space (Merritt, Latorre & King 2015, Merritt & King 2013, Henter et al. 2014).

In any evaluation, one must carefully consider the task, which includes the question that listeners are asked to answer, and how the results are analysed (Wester, Valentini-Botinhao & Henter 2015). It is seldom effective to ask listeners to attend to very specific aspects of speech; see also (Merritt 2016). Instead, most evaluations use broad and non-specific formulations such as "Rate the quality of the following speech samples". Given that nearly all relevant experimental results are based on this type of question, this chapter is restricted to considering the effects of different design choices on generic "quality" or "naturalness". That said, the question asked in a subjective test *is* important and can influence the outcome, all other factors being equal, cf. (Dall, Yamagishi & King 2014). If we were to consider limitations in, e.g., synthetic speech intelligibility instead of naturalness, we would find that system rankings change (King 2014). This shows that design choices can impact different metrics in quite distinct ways.

Objective comparisons have obvious advantages of being fast, cheap, and straightforward, but it is notoriously difficult to devise objective criteria that correlate adequately with human judgements (Hinterleitner 2017). Consider the case of the global variance (GV) of speech parameter trajectories. Naïve synthesis systems typically generate parameter trajectories with much smaller dynamic range – that is, less global variance – than those of natural speech, and the result is perceived as "muffled". Reduced variance is commonly taken as evidence of a loosely-defined issue known as *over-smoothing*. By changing either the model or the generation procedure (Zen, Nose, Yamagishi, Sako, Masuko, Black & Tokuda 2007, Shannon & Byrne 2013)(Shannon 2014, Sec. 6.4.4) to match the global variance observed in training data, perceptual output quality is significantly improved. This has been replicated numerous times, with different synthesisers, datasets, and different techniques for re-instating the variance (Toda & Tokuda 2007, Silén, Helander, Nurminen & Gabbouj 2012, Toda, Muramatsu & Banno 2012, Nose 2016). However, acoustic model likelihood will actually be reduced by these techniques: an objectively inferior model produces perceptually superior output.

Learning an objective measure from existing human judgements is a more promising direction. On unseen data, the best current machine-learning predictors have a Spearman's correlation coefficient around 0.6 between predicted and actual per-stimulus mean opinion scores from previous Blizzard Challenge evaluations (Yoshimura, Henter, Watts, Wester, Yamagishi & Tokuda 2016).

Even the best objective criteria frequently fail to identify meaningful differences between stimuli, instead predicting that most stimuli will be judged as close to the average performance, even though the mean scores assigned by human listeners are distributed over a wide range from bad to good. For this reason, objective criteria such as mel-cepstral distortion or parameter estimation objective function value (e.g., data likelihood) should be reserved for when a very large number of comparisons have to be made, such as when tuning parameters during system development (Kominek, Schultz & Black 2008). When we are interested in what actually works in practice – as in this chapter – there is no substitute for carefully elicited human judgements.

1.3.4 The surrounding model

The accuracy and properties of the surrounding model will affect which issues are audible and identifiable. The standard approach is to start from a low-accuracy model – most commonly, a complete baseline text-to-speech system – in which case the output is a lower bound on the maximum performance achievable: "By construction, we know that it is possible to do at least this well". The study (Watts, Henter, Merritt, Wu & King 2016) is a prime example of this approach, and will be cited extensively in this chapter.

A less common approach, but one that can provide insights into aspects of performance that are not audible in the standard approach (perhaps because they are obscured by other degradations), is to start from a high-accuracy speech model. Since we do not actually have such a model available, the effects of different assumptions and design choices can only be simulated by manipulating parameters extracted from natural speech recordings; these samples are assumed to be random samples from the true (but unavailable) model. (Henter et al. 2014) is the most prominent example of this type of study, in terms of the results discussed in this chapter.

It should be emphasized that conclusions from manipulation-based studies generally only apply in the limit of highly accurate models, and it might be possible for less accurate models to surpass them. As an example, we have seen in section 1.3.3 above that low-accuracy acoustic models designed to inflate the global variance are often perceptually superior to models that maximise the likelihood of the training data. Manipulation-based approaches might also introduce unintended processing artefacts. In the study (Henter et al. 2014), several control conditions were introduced, and the listener scores given to these systems showed that vocoding and duration manipulation on their own could not explain the performance degradations uncovered.

Some investigations consider a wide spectrum of different systems and evaluate both modelled and modified speech stimuli together, including some that are above the typical analysis-synthesis top line. An example that typifies this approach is (Merritt, Latorre & King 2015). Hybrid evaluation approaches also exist: when evaluating acoustic models, it is not uncommon to generate speech parameters based on a highly-accurate duration model, namely oracle durations copied from held-out natural speech, as in (Watts et al. 2016) amongst many others. It is thus not possible to interpret the outcome from such evaluations as either lower or upper bounds, or as an indication of what could be expected for an end application. This caveat is especially relevant for prosody, where durations can have a substantial impact on perception.

1.3.5 The output generation method

Finally, the manner in which the model generates speech parameters can also have a very substantial impact on the perceived properties of the output speech. In principle, natural speech can be seen as samples from an unknown and highly complex "true" statistical model of speech. It may therefore seem compelling to generate speech by sampling from trained models. Unfortunately, this exposes severe issues with most parametric synthesis models, and randomly generated output sounds notoriously poor: speech sound durations are highly idiosyncratic, while acoustics change so randomly and rapidly that the output sounds warbly or bubbly. Only with the very recent WaveNet – which is waveform-level rather than statistical parametric speech synthesis – have sampling-based methods been able to generate good signal quality (van den Oord et al. 2016).

In order to avoid the issues associated with random sampling from poor models, virtually every practical TTS system instead uses a deterministic output generation criterion that returns a carefully curated, identical output each time. For acoustic models, the most widespread criterion is so-called *most likely parameter generation* (MLPG) (Tokuda, Yoshimura, Masuko, Kobayashi & Kitamura 2000).[1] This is based on synthesising output from the most probable output sequence under the current speech model, given the input text.

In practice, the most probable output (or *mode*) is very difficult to estimate from real data, and can be slow or infeasible to compute even for a fitted statistical model. Assumptions are used to circumvent this – for instance, instead of integrating over all possible utterance durations and all possible paths through the hidden state space in an HMM, only a single path through the state space is used. To what degree this choice degrades or improves the output has not been studied in depth. Conveniently, whenever the distribution over output trajectories is Gaussian (which is frequently the case with conventional models and single paths through the state-space during generation), the most likely trajectory is simply the mean of this Gaussian model, which can be computed using the algorithms in (Tokuda et al. 2000). It is well established that predicting and synthesising from the resulting mean trajectory sounds much better than random sampling (Uría, Murray, Renals, Valentini-Botinhao & Bridle 2015, for example). In addition, since estimating means is statistically straightforward, it is possible to estimate the mean trajectory of highly accurate models of speech as well; this was done

[1] MLPG is sometimes read as "maximum likelihood parameter generation", but this is something of a misnomer, since a likelihood denotes the probability assigned to a fixed dataset as the model changes, not the probability of variable data for a fixed model as considered here (Ling & Dai 2012); the same generation principle has also been called "maximum output probability parameter generation" (MOPPG) and "standard parameter generation" (Henter et al. 2014, Ling & Dai 2012, Shannon 2014).

in the study by (Henter et al. 2014), with the finding that the mean of highly accurate models is perceptually inferior to random samples from the same model, at least in the domain of speech parameters derived from STRAIGHT. This shows that there is a very substantial interaction between the modelling assumptions and generation techniques: for poor models, random examples sound worse than the mean, while for accurate models the reverse is true.

In practice, there are other deterministic generation schemes that tend to be subjectively preferred over the raw (approximate) most probable parameter trajectory of a maximum-likelihood fitted acoustic model. These revised output generation procedures are generally based on the idea of compensating for the lower-than-expected global variance (GV) of speech produced by the standard deterministic procedure.

The GV deficiency arises because, somewhat surprisingly, the “most probable” output sequence is not guaranteed to be a typical example of speech. It may actually be far from natural. As stated in section 1.3.4 above, insufficient GV in synthesised speech appears to correlate with poor subjective scores, and the GV can be boosted either by changing the model to yield more appropriate GV under MLPG, or by changing the generation principle instead. The latter usually involves changing the MLPG output prior to playout, which is frequently called *post-processing*, or *postfiltering* after an early post-processing paradigm in speech compression (Ramamoorthy & Jayant 1984). (Note, however, that correlation does not imply causation, and issues with generation methods being GV-deficient may not be the only reason that GV-boosted output frequently is considered perceptually superior.) Some other approaches to this are variance scaling (Silén et al. 2012) and methods incorporating global affine transformations (Toda et al. 2012, Nose 2016, Ling & Dai 2012).

As a final note, the generation methods used also affect what objective evaluation criteria that are seen as most appropriate: whereas maximum likelihood, or similar criteria which assess the accuracy of the predicted speech parameter distribution, are most informative for the case of sampling-based generation, methods that consider only the generated output (such as RMSE or MCD) might be superior for deterministic techniques. Since post-processing often has an adverse effect on standard objective metrics, it is common to apply post-processing only for subjective listening tests, but not when making objective comparisons.

1.4 Shortcomings of Statistical Models for Speech Synthesis

Having described in section 1.3 above a general methodology for attributing degradations to modelling assumptions and other design choices, and for assessing their severity, this section will present empirical findings regarding:

1. Duration modelling, in section 1.4.2
2. Machine-learning paradigm (acoustic regression model), in section 1.4.3
3. Across-context averaging, in section 1.4.4
4. Distribution and dependence assumptions, in section 1.4.5

5. Joint or separate stream modelling, in section 1.4.6
6. Temporal modelling, in section 1.4.7
7. Optionality, in section 1.4.8

We begin by briefly introducing the two main comparative studies upon which the subsequent discussion is based.

1.4.1 Key comparative studies

(Watts et al. 2016) and (Henter et al. 2014) are both side-by-side MUSHRA tests that analyse the relative impacts of a chain of different modelling assumptions and design choices. The former focussed on modelled speech (a low-naturalness operating point) and the latter on manipulated speech (a high-naturalness operating point).

The study in (Watts et al. 2016) compared a number of different text-to-speech systems designed to interpolate between, atone end, a state-of-the-art decision-tree-based speech synthesis system (HTS, (Zen, Nose, Yamagishi, Sako, Masuko, Black & Tokuda 2007)) and, at the other end, a recent deep-neural-network-based synthesiser (Merlin, (Wu, Watts & King 2016)) using feedforward DNNs. The systems were all trained on the same database and used the same vocoder (STRAIGHT). All used oracle durations from held-out natural speech, making these studies an investigation into the performance of different acoustic modelling techniques.

The main aim of (Watts et al. 2016) was to identify the key factors that contribute to the empirically-observed improvement in performance of newer DNN-based TTS systems over established decision-tree-based synthesisers like HTS. The implementational differences between systems from the two paradigms are not limited to the machine-learning technique, but include many additional choices; those examined were:

1. Regression model: decision trees (DT) or feedforward deep neural networks (NN) (section 1.4.3).
2. Temporal granularity: piece-wise constant for each sub-phone state, or changing every frame (section 1.4.7).
3. Stream modelling: each stream of speech parameters can be predicted by a separate regression model or all can be predicted together by a single, joint model (section 1.4.6).
4. Variance model: the variance used during generation can be predicted by the regression model or can be a global constant (section 1.4.5).
5. Duration-dependent input features: whether duration is used by the acoustic regression model (section 1.4.2).
6. Trajectory enhancement method: global variance modelling (GV) (Toda & Tokuda 2007) or regular mel-cepstral domain postfiltering for formant enhancement (PF) (Yoshimura, Tokuda, Masuko, Kobayashi & Kitamura 2005). (sections 1.3.5 and 1.4.4).

The configurations of the different TTS systems built, and specifically their differences in terms of the above-mentioned aspects, are listed in Table 1.1.

ID	Model	Resolution	Streams	Variance	Dur. dep.	Enhancement
V	-	-	-	-	-	-
D1	DT	state	separate	local	no	GV
D2	DT	state	separate	local	no	PF
N1	NN	state	separate	local	no	PF
N2	NN	state	separate	global	no	PF
N3	NN	state	joint	global	no	PF
N4	NN	frame	separate	global	no	PF
N5	NN	frame	joint	global	no	PF
N6	NN	frame	joint	global	yes	PF

Table 1.1 An overview of the TTS systems compared in (Watts et al. 2016), showing their IDs and the factors that were successively altered in order to step from HTS ("D1") to Merlin ("N6"). More information about the different factors is provided in the text and in (Watts et al. 2016). "V" is analysis-synthesised speech acting as a top line reference in the listening test.

After building eight different text-to-speech systems as listed in Table 1.1, output from these systems plus vocoded natural speech were compared in a MUSHRA test. In the test, 20 native, paid listeners each scored parallel system output on 20 phonetically-balanced sentences (selected for each participant in a balanced manner from a pool of 70), for a total of 400 parallel ratings. The results of the test are illustrated in figure 1.2. The stimuli and ratings are freely and permanently available online.[2] Upon applying double-sided Wilcoxon signed-rank tests to all system pairs, with a Holm-Bonferroni correction (Holm 1979) to keep the familywise error rate below $\alpha = 0.05$, the different systems studied separated into five distinct sets, such that all between-set comparisons were statistically significant, whereas all within-set comparisons were not. The sets are delimited by dotted vertical lines in figure 1.2. From these results, it seemed that the major gains in synthesis performance between systems "D1" and "N6" coincided with the switch from decision trees to neural networks in the regression model and with the change from state-level to frame-level time granularity. Adding duration-derived features also made a significant difference, although the effect size was smaller; it is unclear if this would apply when using predicted rather than oracle durations. While GV is generally considered superior to regular formant-enhancement postfiltering, that difference was by comparison not so large as to be significant in this investigation, though the difference between "D1" and "D2" was judged as significant if a per-subject score normalisation was introduced (Watts et al. 2016). The implications of the study findings are discussed more in-depth in section 1.4.3 onwards.

The second main study (Henter et al. 2014) evaluated only natural and manipulated

[2] doi:10.7488/ds/1316

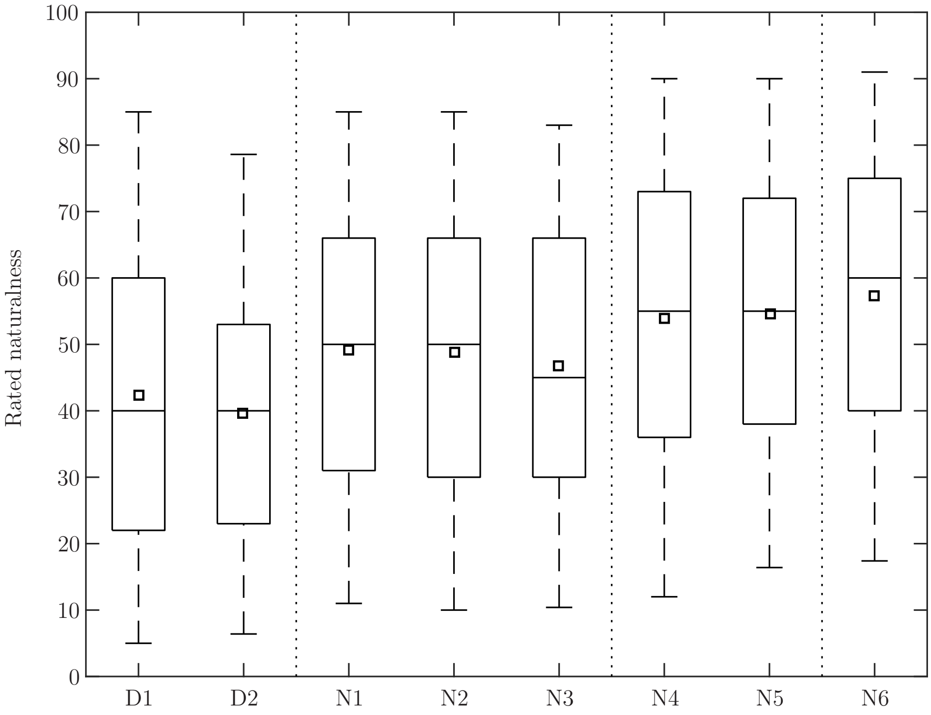
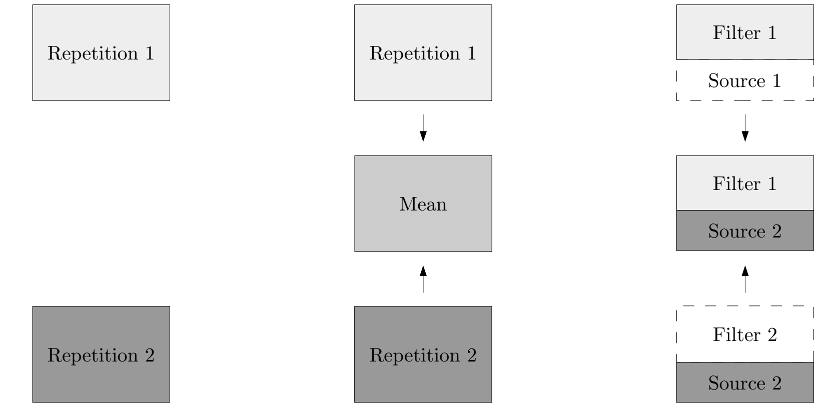


Figure 1.2 Box plot of aggregate listener naturalness ratings from the MUSHRA test in (Watts et al. 2016). Condition labels are as in table 1.1; "V" is omitted as it was rated at 100 (maximally natural) more than 95% of the time. For the boxes, middle lines show medians, box edges are at quartiles, while whiskers extend to cover all but 5% of data on either side. Squares denote the mean rating. Dotted lines separate systems into sets, where systems within a set exhibited no statistically significant differences in rating, while all cross-set differences were statistically significant.

speech, making it an investigation of the upper limits placed on naturalness by a number of modelling assumptions and design choices in speech synthesis, in the context of a highly accurate model. The central innovation was to use a carefully purpose-recorded database of repeated speech, called the Repeated Harvard Sentence Prompts (REHASP) corpus version 0.5, where the same sentence prompt was read aloud multiple times by the same speaker in identical conditions. Each recording can then be seen as a statistically independent sample from the same true speech distribution for that particular sentence. This database is freely available.[3]

By applying dynamic time warping, all repetitions of the same prompt were also made to have the same timings. (The mathematical interpretation of this is that, for each frame, all time-warped repetitions were in the same state in a left-right state-space model at that frame. Due to how HMMs are defined, different parameter trajectories can then be treated as conditionally independent.) This allowed the different repetitions to

[3] doi:10.7488/ds/39



(a) Aligned repetitions (b) Mean speech (c) Chimeric speech

Figure 1.3 Manipulating repeated speech. After obtaining time-aligned independent readings (repetitions) of the same prompt, shown in 1.3(a), the aligned speech-parameter matrices can be blended into an estimate of the true average speech, as in 1.3(b), or stitched together to form chimeric speech that represents randomly-sampled speech output from a highly accurate model with certain independence assumptions, such as conditional source-filter independence in 1.3(c).

be combined in various ways, either creating *chimeric speech* by taking the trajectories of different features from different, independent repetitions, or blending all repetitions into one grand average; see figure 1.3. The result was manipulated speech stimuli approximating the output of highly accurate trajectory models, under various conditional-independence assumptions (no independence versus independence between source and filter, parameter streams, and filter coefficients) and generation methods (random sampling or taking the mean). Table 1.2 details how the most relevant conditions from the study were constructed, as well as their interpretation.

30 native, paid listeners compared the naturalness of different manipulated speech examples for the same prompt in a MUSHRA test. The results of this test, based on a total of 549 parallel ratings, are summarised in figure 1.4. Bonferroni-corrected pairwise *t*-tests found all system pairs to be significantly different at the 0.01 level, except (SF, SI) and (SI, M). The results thus lend some insight into how much the tested assumptions might limit speech synthesiser naturalness, as discussed in the remainder of this chapter. It should, however, be pointed out that the results, strictly speaking, only are known to be valid for the specific speech parameterisation used in the study, which was based on legacy STRAIGHT with a mel-cepstrum representation of the STRAIGHT spectrogram.

1.4.2 Duration modelling

A probabilistic model of speech parameter sequences is usually factored into a duration model and an acoustic model, where durations typically have to be predicted first dur-

Condition		Parameter source repetitions				Interpretation		
ID	Description	Dur.	LF0	BAPs	MCEPs	Domain	Model	Generation
N	Natural waveform	-	-	-	-	Waveform	True	Sampling
V	Vocoded	a	a	a	a	Param.	True	Sampling
D	Time-warped	b	a	a	a	Param.	$\mathcal{M}_{\mathrm{D}}$	Sampling
SF	Source/filter indep.	b	a	a	c	Param.	$\mathcal{M}_{\mathrm{SF}}$	Sampling
SI	All streams indep.	b	a	d	c	Param.	$\mathcal{M}_{\mathrm{SI}}$	Sampling
I	MCEPs indep.	b	a	a	$*$	Param.	$\mathcal{M}_{\mathrm{I}}$	Sampling
M	MCEPs averaged	b	a	a	mean	Param.	$\mathcal{M}_{\mathrm{D}}$–$\mathcal{M}_{\mathrm{I}}$	Mean

Table 1.2 An overview of the main conditions (speech manipulations) investigated in (Henter et al. 2014), showing their ID and description, where different parameter trajectories were sourced for each manipulation and how they are to be interpreted. For the parameter trajectory sources, each different letter represents a different source repetition that was used; "$*$" means that each coefficient trajectory was taken from a different repetition in the database, while "mean" is an average over all repetitions. The interpretation-related fields distinguish the domain of the synthesis (waveform or speech parameters), the statistical model used and the output generation method.

ing synthesis. We will therefore start by briefly discussing the duration model, before acoustic modelling.

While duration is an important component of synthesising speech there have been relatively few studies that have tried to isolate the effects of different design choices in modelling duration, compared to the extensive literature on acoustic modelling. The simplest possible duration model is to assume that all (sub-state) durations are independent and given by a discrete, no-skip, left-right Markov chain. This was a common model in early decision-tree-based synthesis. In this model, advancing to the next state is essentially decided by a coin toss (a Bernoulli random variable), so that individual state durations implicitly follow a geometric distribution. This model is a poor fit with actual durations in state alignments on training data (for one thing, the most probable duration is always a single frame), but is acceptable for synthesis as long as the expected duration is used at generation time, noting that HMM training (maximum likelihood parameter estimation) of duration parameters in this case reduces to matching the expected duration of the model with the mean durations observed in the training data.

(Zen, Tokuda, Masuko, Kobayashi & Kitamura 2007) proposed improving the duration model by replacing the Markov model over the state space (and the implicit geometric duration distributions it entails) with a semi-Markov model, which is memoryless given the current state *and* how many time steps the process has spent in the current state. This allows state durations to follow any discrete-valued stochastic distribution. The resulting construction is known as a *hidden semi-Markov model* (HSMM). In practice, it is often assumed that durations follow a Gaussian distribution, even though

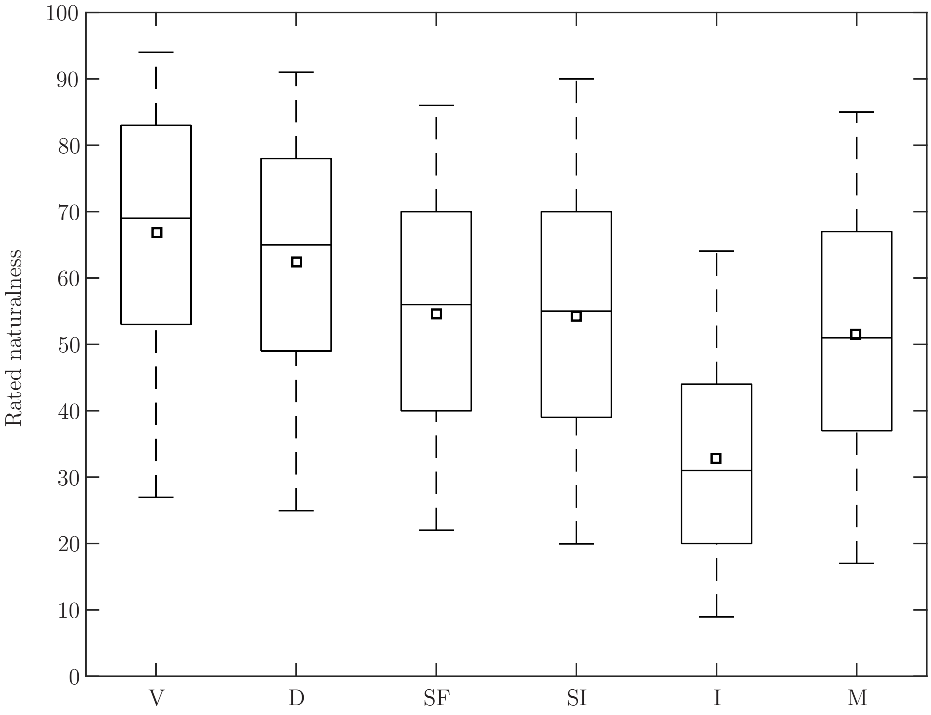


Figure 1.4 Box plot of aggregate listener naturalness ratings from the MUSHRA test in (Henter et al. 2014). Condition labels are as in table 1.2; "N" is omitted as it always was rated at 100 (completely natural). For the boxes, middle lines show medians, box edges are at quartiles, while whiskers extend to cover all but 5% of data on either side. Squares denote the mean rating. All system pairs showed significant differences in rating except (SF, SI) and (SI, M).

this distribution is continuous-valued (not discrete) and can attain negative values. This choice was seen to improve the subjective quality of synthesised speech in the small listening test in (Zen, Tokuda, Masuko, Kobayashi & Kitamura 2007), and is incorporated as standard in HTS, though the relative importance of the improvement has not been well studied in relation to other problem sources. HSMMs were also recently used for EM-based realignment in a feedforward DNN-based speech synthesiser (Tokuda, Hashimoto, Oura & Nankaku 2016). Either way, since the change does not directly alter the fundamentals of how durations are predicted – predictions are still based on the mean duration of the state duration, which still is estimated as the sample mean over aligned phones allocated to the relevant decision tree node – any improvements an HSMM brings over an HMM are likely to be due to improved alignments indirectly benefiting both duration models and acoustic models, by more accurately associating frames with a suitable state in a contextual phone.

A substantially different approach to speech synthesis duration modelling was recently suggested by (Ronanki, Watts, King & Henter 2016). They replaced the parametric Gaussian distributions of traditional HSMMs by a non-parametric distribution predicted by the same recurrent neural network that also predicted acoustic parameters, and used distribution quantiles (rather than the mean) for output generation. It is, however, too early to tell if this approach will bring subjective improvements in the long run.

1.4.3 The machine learning paradigm

The machine learning paradigm used to predict speech properties from text-extracted linguistic features is the other major factor that affects duration modelling. This area has seen a major paradigm shift in recent years, with decision trees being replaced by deep and recurrent learning techniques. For acoustic modelling, specifically, it has generally been found (Wang, Takaki & Yamagishi 2016*a*) that decision-tree regression is inferior to deep feedforward neural networks (Zen, Senior & Schuster 2013, Zen & Senior 2014, Hashimoto, Oura, Nankaku & Tokuda 2015, Watts et al. 2016), which in turn tend to be outperformed by recurrent neural network techniques (Fan, Qian, Xie & Soong 2014, Zen & Sak 2015), of which the long short-term memory (Hochreiter & Schmidhuber 1997) has received the most attention. The study by (Watts et al. 2016) is especially elucidating, since it compared the relative impact of many different aspects that distinguish conventional decision-tree-based systems from recent feedforward-DNN-based TTS, and found that the change in machine learning paradigm yielded one of the most notable improvements (both in numerical size and statistical significance) observed in their quantitative MUSHRA listening test.

While feedforward or recurrent neural networks have become central to high-quality speech synthesisers in many applications, there are alternative regression techniques as well, such as random forests (Black & Muthukumar 2015). They were also able to show improved modelling accuracy over only using a single decision tree for regression.

1.4.4 Cross-context averaging

This issue arises when the synthesis approach fails to distinguish frames that, conditioned on the input text, should be realised as acoustically distinct, thus treating training examples as interchangeable when they are not. Cross-context averaging has been found to be particularly harmful to synthesis quality (Merritt, Latorre & King 2015, Merritt 2016). It is easy to see that this situation can occur in decision-tree-based approaches, which rely on clustering training data frames together and assigning them to leaves in a tree: only aggregate properties of these clusters are used in synthesis.

Cross-context averaging is not unique to decision trees, however. A similar effect can arise with any machine learning paradigm, in cases where the available input features (linguistic or otherwise) fail to distinguish speech contexts that should be given acoustically distinct realisations, for instance by not properly separating between stressed and unstressed instances of the same word. Mathematically, if there is structured variation that cannot be predicted due to conflating and averaging speech across contexts, this variation will instead be absorbed by the model: the noise term in a Gaussian model. Since the most likely output is the mean (which ignores the noise term) cross-context averaging is likely contribute to the deficient GV of generated parameter trajectories. At the same time the over-inflated noise term is also likely to contribute to the unappealing, noisy and unstructured behaviour of random sampling from such a model.

It was observed that scaling up the variance of synthesised speech to match the training data GV did not completely undo the detrimental effects of the averaging (Merritt, Latorre & King 2015), presumably because the model is incapable of recreating the missing contextual distinctions. However, there are methods to mitigate cross-context averaging, often referred to as *rich-context models*. Several studies support the conclusion that rich context modelling leads to perceptually superior synthesised speech, both in conventional decision-tree-based speech synthesis (Yan, Qian & Soong 2009, Merritt, Yamagishi, Wu, Watts & King 2015) and in hybrid synthesis approaches (Merritt et al. 2016).

A typical rich-context model will distinguish all possible quinphones centred on the current phone. This is not to say that contextual information outside this window cannot be informative as well. Wu et al. (Wu & King 2016, Wu, Valentini-Botinhao, Watts & King 2015) augmented the standard linguistic features with bottleneck-DNN-derived features that summarise the acoustically most salient information contained in the input features of surrounding frames. Synthesis with the augmented features of 23 surrounding frames produced synthetic speech that was subjectively preferred. It is likely that improvements provided by recurrent neural networks (Fan et al. 2014, Zen & Sak 2015) are also due to better propagation of linguistic feature information across frames.

Surprisingly, rich linguistic context alone may never be sufficient for achieving truly convincing speech output, even with utterance-length contextual information, and a model close to true speech. This was explored by (Henter et al. 2014), as outlined in section 1.4.1, performed frame-wise alignment using dynamic time warping of 40 recordings of the same single-sentence prompt, read aloud by a single speaker. By averaging the 40 recordings (frames) for each warped time instant, a stimulus “M” was obtained that closely approximates the conditional mean of the “true” speech model, given the perfectly-matched utterance-wide linguistic context of the frame. Nevertheless, as seen in figure 1.4, the result was significantly inferior in naturalness to both analysis-synthesised speech (“V”) and vocoded speech with time-warped durations (“D”). We can conclude that the operation of averaging – even when performed over exceptionally comparable acoustic frames whose entire linguistic context is identical – limits the naturalness that can be achieved.

1.4.5 Distribution and dependence assumptions

Most issues discussed so far revolve around how the central tendency (i.e., mean) of the speech is described by the regression model. However, the noise (or covariance) model, used for describing the distribution of deviations from the regression model prediction within each time frame, can also have an impact, particularly for stochastic output generation methods.

The simplest possible description of acoustic frame properties is not to use a model but to minimise a distance measure. In practice, the mean squared prediction error is used almost exclusively. Mathematically, this corresponds to a context-independent isotropic Gaussian with all output dimensions having the same variance. One step up in complexity is a Gaussian with diagonal covariance matrix, so that each dimension has a different (still context-independent, i.e., global) variance. These two setups are commonplace in recent deep learning-based speech synthesis, with the latter being the standard in Merlin (Wu et al. 2016). The next step is to allow variances to depend on context, an to predict them with the regression model. This is standard in decision-tree based synthesisers such as HTS (Zen, Nose, Yamagishi, Sako, Masuko, Black & Tokuda 2007). In a comparative study (Watts et al. 2016), models with global versus context-dependent variance ("N2" vs. "N1") were not found to sound very different.

A further extension is to consider full (i.e., non-diagonal) covariance matrices, though this can easily lead to a model with a very large number of degrees of freedom, susceptible to overfitting. So far, models with full covariance matrices have mostly appeared in decision-tree-based speech synthesisers, for example semi-tied covariance matrices (Gales 1999). In any case, the effect on deterministically generated speech output is likely to be subtle, with a minor improvement claimed in (Zen, Nose, Yamagishi, Sako, Masuko, Black & Tokuda 2007). Any improvement comes at greatly increased computational complexity of synthesis.

By relaxing the Gaussian assumption, the distribution of speech frame features, and not just their means and variances, can be described and predicted more accurately. In these cases, the most probable output, if used for synthesis, may no longer coincide with the mean of the distribution, so the limitations identified for condition "M" in (Henter et al. 2014) may not apply. One example would be real-valued neural autoregressive density-estimators (RNADEs) (Uría, Murray & Larochelle 2014), which let the distribution of feature components depend on the outcome values of preceding components through a neural network. When applied to TTS by (Uría et al. 2015), RNADEs were seen to substantially improve held-out data likelihood, and produced deterministically generated speech that was preferred by listeners.

Another non-Gaussian approach to describing speech feature distributions within a time frame is offered by mixture models. In particular, Gaussian mixture models whose parameters are predicted by a deep neural network (an instance of the mixture density networks, or MDNs, described in (Bishop 1994)) have given modest improvements (Zen & Senior 2014, Wang, Xu & Xu 2016, Henter, Ronanki, Watts, Wester, Wu & King 2016). For computational convenience, these methods usually only consider the most massive mixture component when generating output. The benefit of the remaining mixture components may simply be to absorb difficult-to-explain training data that would otherwise 'pollute' the most massive component (Henter et al. 2016).

If we consider generation based on random sampling, the importance of covariance modelling increases dramatically. (Uría et al. 2015) uncovered a substantial preference in favour of samples drawn from an RNADE model compared to samples drawn from a similar model with no explicit modelling of conditional dependences between speech features.

1.4.6 Separate or joint modelling of parameters

Related to dependence assumptions is the topic of between-feature dependence: joint or separate modelling of speech parameter streams, where “streams” are subsets of the acoustic features (sometimes also including durations) that are believed to behave similarly in terms of modelling. STRAIGHT features are usually partitioned into three streams, namely (log) F0, the aperiodicity coefficients, and the spectral envelope coefficients (mceps). Decision-tree-based synthesis, in particular, can benefit from using separate regression trees for these different feature streams.

In the deep learning paradigm, speech recognition performance has been seen to improve through *multi-task learning*, where the NN is trained to simultaneously solve an additional, related prediction task (Qian, Yin, You & Yu 2015) and this idea has also been applied to synthesis (Wu et al. 2015).

The conjecture is that lower DNN layers learn to capture more universal and generalisable structure from the data: the additional task acts as a regulariser. So – unlike decision trees – it should be beneficial to predict all output streams using a single, joint NN. (Chen, Chen, Xu & Yu 2015) found that using a single feedforward DNN improved the subjective MOS compared to separate DNNs for each stream. (Watts et al. 2016) also considered this distinction, but did not uncover any noticeable differences.

While separate versus joint stream modelling can influence how cross-stream dependencies are modelled, it is *not* the same as a statistical independence assumption between output features. As a counterexample, most models that predict feature streams using a joint model still assume diagonal-covariance Gaussian distributions. The impact of conditional independence assumptions between streams was investigated by (Henter et al. 2014) who found that speech randomly sampled from models that assume parameter streams to be conditionally independent, but otherwise are highly accurate and internally consistent within streams, is subjectively inferior to speech samples from highly-accurate models that also account for cross-stream dependences (conditions “V” vs. “SI” in the study).

1.4.7 Trajectory modelling

Thus far, the discussion of acoustic feature modelling has primarily covered regression techniques, distribution assumptions, and feature dependences within single, isolated time frames. But speech is a stochastic time series, and so it is also necessary to model the temporal evolution of acoustic parameters. This modelling encompasses two parts: assumptions about the stochastic distributions of speech parameter trajectories for a given text, and how the properties of these distributions are made to depend on the text input across time (and in particular the temporal resolution of this dependence).

Many statistical speech synthesis systems account for time-dependence in similar way to automatic speech recognition, by modelling not only so-called ‘static’ frame-wise features, but also the local differences (deltas) and second-order finite differences (delta-deltas) of the frame-wise features: collectively known as *dynamic features*. In a Gaussian model of parameter trajectories over time, the use of dynamic features means that the precision matrix must have a band-diagonal structure, and time dependence is restricted to simple, linear correlations. Statistically, the result is a product-of-experts model, where the most likely output is generated as a compromise between soft constraints on the static and dynamic properties of the trajectory (Shannon 2014, Sec. 3.3.2). Practically, the consequence is a smoothed output sequence, compared to the piecewise-constant output if only a model of the 'statics'.

The relative impact of temporal smoothing on natural speech parameter trajectories was been investigated in (Merritt & King 2013) where it was found that the temporal averaging produced by such smoothing had a much smaller effect on naturalness than the cross-context averaging discussed in section 1.4.4 above. This is consistent with (Zhang, Tao, Jia & Wang 2008), who found overly smooth temporal trajectories less problematic than overly smooth spectra.

A majority of synthesisers ignore the deterministic relationship between static and dynamic feature values during training, only taking them into account during output generation: so, the output is generated from a different model than that created during training (Zen, Tokuda & Kitamura 2007, Shannon, Zen & Byrne 2011). A complementary perspective is that the normalisation constant used during training is incorrect, since it accounts for combinations of static and dynamic feature values that are simply impossible. A consequence of the mismatch is that the trained trajectory model severely underestimates the statistical variation possible in natural speech parameter trajectories, and therefore tends to assign pathologically small probabilities to held-out speech utterances (Shannon et al. 2011). Using matched and properly normalised models during both training and synthesis may be perceptually superior to the conventional, mismatched approach (Zen, Tokuda & Kitamura 2007).

In decision-tree based models, such as HTS, each context-dependent model comprises a small number of states, and so the statistics of the generated output remain the same for several consecutive frames. Systems exist, however, for which the statistical properties of every frame can be different. A common mechanism is to provide the current position *within* state or phone as an input to the acoustic predictor. This is straightforward, even standard, in neural network-based acoustic models like Merlin, but the same type of feature can with some additional effort also be integrated into synthesisers using Gaussian process regression (Koriyama, Nose & Kobayashi 2014) or decision-trees. The canonical example of the latter is Clustergen (Black 2006), where the time resolution is improved by subdividing states (decision tree leaves) by thresholding the position indicator.

For neural networks, it is easy to compare parallel synthesisers that either include or omit these frame-level positional features. The MUSHRA test in (Watts et al. 2016) found that their inclusion gave one of the most substantial quality improvements, making this a key difference between decision tree approaches such as HTS (Zen, Nose, Yamagishi, Sako, Masuko, Black & Tokuda 2007), and deep learning systems such as Merlin (Wu et al. 2016). In contrast, the 10-subject MOS test in (Tokuda et al. 2016) found very similar performance between a DNN-based synthesiser with frame-level granularity using oracle durations, and a feedforward DNN with state-level granularity and durations predicted from a HSMM duration model, where the DNN-HSMM system had been trained using the generalised EM-algorithm. It is not straightforward to reconcile these two, seemingly conflicting findings.

Of course, recurrent neural networks, regardless of the input features, are able to learn to model fine-grained positional information, using the internal network state.

1.4.8 Optionality

The findings in (Henter et al. 2014) suggest that, in order to be truly acoustically convincing, speech synthesis needs to move beyond deterministic output generation methods that merely attempt to produce average speech. Many systems already perform post-processing of the generated mean speech to better match the global variance of natural speech, which does increase subjective output naturalness (Silén et al. 2012, Toda et al. 2012, Nose 2016). But, the effects of such processing have not been studied at a high-accuracy operating point, where it seems less likely to be effective. A more radical change would be to generate speech based on something other than the expected value, for instance by sampling, as discussed in section 1.3.5, something that we are just starting to see in very recent work.[4]

A third alternative is to allow *optionality* in the speech realisation, which here is taken to mean more than a simple left-right model. A concrete example would be alternating emphasis, driven by information not available within the text (Ribeiro et al. 2015). Adding 'external' inputs to distinguish between different possible realisations can be achieved with a multiple-regression HMM (MR-HMM) (Takashi, Tachibana & Kobayashi 2009). Among the three changes that stuck out in (Watts et al. 2016) as significant improvements, one was that of using expanded duration-dependent input features that allow for easy representation of optionality.

1.5 Conclusions

We have seen that current approaches to both vocoding and statistical modelling limit the naturalness of contemporary parametric text-to-speech systems in a variety of ways. A natural question is then – given what we have learned about these limitations from empirical data – what are the most appealing research problems to pursue, in order to create improved parametric speech synthesisers?

For vocoding, sinusoidal signal representations appear capable of providing a low-level representation of the speech signal that maintains very high quality. The question remains how to use this signal-fitting capacity for building a high-quality vocoder that allows faithful reconstruction of all perceived characteristics of the speech signal in the absence of any statistical modelling.

As more and more parts of speech synthesisers have been replaced by neural networks optimised via stochastic gradient descent (cf. (Takaki et al. 2015)), the long-term goal of integrating aspects of the signal processing into the learned parts of the model

[4] For instance, authors' pre-prints not peer-reviewed at the time of writing.

has come to the forefront. The most extreme version of this agenda is statistical modelling in the raw waveform domain, which does not require a vocoder and may enable joint end-to-end optimisation of both signal processing and statistical modelling. This line of research recently saw a breakthrough in the form of the WaveNet paper from Google DeepMind (van den Oord et al. 2016). A major downside, however, is the computational cost of the present WaveNet synthesiser, which simply is infeasible for practical applications. Reducing the computational load of these approaches whilst maintaining naturalness will no doubt be an important area of future research. In the meantime, non-parametric approaches to waveform generation continue to achieve the greatest segmental quality achievable in TTS applications, and it is not unreasonable to surmise that signal generators based on recorded speech will dominate applications for some time still.

In modelling, the advent of synthesisers based on deep and recurrent neural networks opened up a path to circumvent several long-standing limitations of decision-tree-based synthesis approaches. Most importantly, the use of neural networks reduced the amount of inappropriate conflation, and thus inappropriate averaging, performed by the learner: this includes both averaging across linguistic contexts, as well as along temporal positions, i.e., frame vs. state-level granularity. In all likelihood, however, future research will identify new setups and approaches that better generalise from these contexts to the distribution of acoustic parameters or – for waveform-level modelling – the joint distribution of audio signal sample values.

There is another side to the conflation coin as well: instead of the all-or-nothing, binary averaging that decision tree methods perform, wherein datapoints in different leaves are treated entirely independently, deep learners might improve their models by learning from related but not directly relevant material, e.g, improve their modelling of one speech sound with the help of data from another, or better model one speaker by using additional speakers in the training data, as observed in (Fan, Qian, Soong & He 2015) and also pursued in (van den Oord et al. 2016). The argument is that the hierarchical design of deep learners allows useful information in otherwise unrelated linguistic contexts be extracted and processed into an abstracted form that is useful across context boundaries, a phenomenon dubbed the *blessing of abstraction* (Tenenbaum, Kemp, Griffiths & Goodman 2011). Historically, advances in speech synthesis have fed off the exponential growth of speech corpora sizes and computational resources seen in the last decades, and deep learning appears to benefit disproportionately much in the limit of very large amounts of data. The WaveNet paper, trained on 44 hours of data from more than 100 different speakers, is no exception to this rule. Pursuing methods that can adequately generalise from the vast amounts of unlabelled, multi-speaker, spontaneous speech material available all around us, even if it is not always directly relevant to the speech we want to synthesise, therefore appears to be another promising direction for future research.

In terms of generating output from distributions, we have learned that predicting the mean speech parameter sequence is not the route to naturalness, at least for speech parameterisations similar to STRAIGHT. From a theoretical perspective, only generation by sampling is – essentially by definition – certain to be able to achieve completely natural speech. Reaching high naturalness with sampled speech, however, places substantial demands on model accuracy: it is no longer sufficient to represent streams and individual parameters independently, but their dependences and collective behaviour must be accounted for in order to rise above the limitations identified in (Henter et al. 2014). In other words, more than one possible outcome must be represented well.

The WaveNet paper (van den Oord et al. 2016) is probably the first published example of a synthesiser where random sampling has produced competitive-sounding speech, though it operates in the waveform domain. Parametric speech synthesis is also likely to see additional efforts to perform accurate dependence modelling, but whether or not sampled parametric speech will come to surpass deterministically generated speech is hard to tell. This is especially true given that the limitations of mean-based generation do not necessarily preclude the existence of other deterministic methods capable of generating completely natural speech, if a suitable model is provided. For instance, it is presently unknown whether or not mean speech with GV-compensation (following, e.g., the methodology of (Nose 2016)) suffers the same upper limit as mean speech without such compensation. It is also not known what the upper limits on most likely parameter generation are – or even whether or not that principle is subjectively preferable to mean speech output generation – in models where the mean and mode do not coincide. These might be topics of future research.

The ultimate goal in speech technology research is not only to create synthesisers that sound natural, but enable technology that is natural to use. Considering the various flaws of contemporary speech synthesisers, the greatest practical issue may not be their segmental quality (which, WaveNet aside, can be made quite convincing in applications using hybrid synthesis from large speech corpora and by targetting the recorded prompts to the specific application domain, as evidenced by systems like Apple's Siri, Google Now, Microsoft Cortana, etc.), nor their intelligibility (which tends to be at ceiling in quiet conditions), but their awkward prosody and their ignorance of the communicative nature of speech and dialogue. It is often speculated, for instance by invoking references to the so called "uncanny valley" (Mori 1970, Moore 2012), that the adoption of TTS for practical tasks is limited not by segmental quality, but by the perceived unpleasantness of TTS systems. Furthermore, it is widely surmised that poor text and dialog/context understanding is a key bottleneck, and that more appropriate prosody and communication ability would be possible with improved linguistic features that better represent semantics and pragmatics.

An alternative, more practical approach may be to pursue better natural language processing, and from it derive more advanced and potentially more informative features for speech synthesis, just to see how much this can improve text-to-speech synthesis. Similar to the situation in speech processing, text and language processing have also seen substantial improvements from the adoption of deep learning, especially given the ease of acquiring truly gigantic text databases; NLP methods may be trained on much more text than any one human may read in their lifetime. If text and language processing continue to improve at the present rate, future developments will provide many interesting candidate techniques for integration into TTS. As a case in point of such technology transfer, one may consider NLP methods like *word2vec* (Mikolov, Chen, Corrado & Dean 2013) and related vector-space representations of text, or text and speech jointly (Rendel, Fernandez, Hoory & Ramabhadran 2016, Wang, Takaki & Yamagishi 2016*b*, Watts, Yamagishi & King 2010, Merritt, Yamagishi, Wu, Watts & King 2015, for example).

Whether through advanced features, through improved statistical sequence modelling, or from some other insight out of left field, accurate modelling and generation of prosody might very well be the final frontier in making synthetic speech pleasant, or at least palatable, to humans, in the context of an application. This might, in turn, be necessary for realising the transformative potential of TTS in places where speaking machines are scarcely more than a novelty, by finally providing humans with technological tools that are not merely powerful, but also natural to use.

Bibliography

Agiomyrgiannakis, Y. & Stylianou, Y. (2009), ‘Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech’, *IEEE Trans. Audio, Speech, Language Process.* **17**(4), 775–786.

Babacan, O., Drugman, T., Raitio, T., Erro, D. & Dutoit, T. (2014), Parametric representation for singing voice synthesis: A comparative evaluation, *in* ‘Proc. ICASSP’, pp. 2564–2568.

Bishop, C. (1994), Mixture density networks, *in* ‘Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University’.

Black, A. W. (2006), CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling, *in* ‘Proc. Interspeech’, Vol. 7, Dresden, Germany, pp. 1762–1765.

Black, A. W. & Muthukumar, P. K. (2015), Random forests for statistical speech synthesis, *in* ‘Proc. Interspeech’, Vol. 16, Dresden, Germany, pp. 1211–1215.

Bollepalli, B., Raitio, T. & Alku, P. (2013), Effect of MPEG audio compression on HMM-based speech synthesis, *in* ‘Proc. Interspeech’, Vol. 14, Lyon, France, pp. 1062–1066.

Borg, I. & Groenen, P. J. F. (2005), *Modern Multidimensional Scaling: Theory and Applications*, 2 edn, Springer Science+Business Media.

Cabral, J. P., Renals, S., Richmond, K. & Yamagishi, J. (2007), Towards an improved modeling of the glottal source in statistical parametric speech synthesis, *in* ‘Proc. ISCA SSW’, Vol. 6, Bonn, Germany, pp. 113–118.

Chen, B., Chen, Z., Xu, J. & Yu, K. (2015), An investigation of context clustering for statistical speech synthesis with deep neural network, *in* ‘Proc. Interspeech’, Vol. 16, Dresden, Germany, pp. 2212–2216.

Cooke, M., Mayo, C. & Valentini-Botinhao, C. (2013), Intelligibility-enhancing speech modifications: the Hurricane Challenge, *in* ‘Proc. Interspeech’, Lyon, France, pp. 3552–3556.

Creer, S., Cunningham, S., Green, P. & Yamagishi, J. (2013), ‘Building personalised synthetic voices for individuals with severe speech impairment’, *Comput. Speech Lang.* **27**(6), 1178–1193.

Dall, R., Yamagishi, J. & King, S. (2014), Rating naturalness in speech synthesis: The effect of style and expectation, *in* ‘Proc. Speech Prosody’, Vol. 7, Dublin, Ireland.

Degottex, G. & Erro, D. (2014), ‘A uniform phase representation for the harmonic model in speech synthesis applications’, *EURASIP J. Audio Spee.* **2014**(1), 38.

Degottex, G., Lanchantin, P., Roebel, A. & Rodet, X. (2013), ‘Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis’, *Speech Comm.* **55**(2), 278–294.

Degottex, G. & Stylianou, Y. (2013), ‘Analysis and synthesis of speech using an adaptive full-band harmonic model’, *IEEE Trans. Audio, Speech, Language Process.* **21**(10), 2085–2095.

Drugman, T., Kane, J. & Gobl, C. (2014), ‘Data-driven detection and analysis of the patterns of creaky voice’, *Comput. Speech Lang.* **28**(5), 1233–1253.

El-Jaroudi, A. & Makhoul, J. (1991), ‘Discrete all-pole modeling’, *IEEE Trans. Signal Process.* **39**(2), 411–423.

Espic, F., Valentini-Botinhao, C., Wu, Z. & King, S. (2016), ‘Waveform generation based on signal reshaping for statistical parametric speech synthesis’, *Proc. Interspeech* **17**, 2263–2267.

Fan, Y., Qian, Y., Soong, F. K. & He, L. (2015), Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis, *in* ‘Proc. ICASSP’, Vol. 40, Brisbane, Australia, pp. 4475–4479.

Fan, Y., Qian, Y., Xie, F.-L. & Soong, F. K. (2014), TTS synthesis with bidirectional LSTM based recurrent neural networks, *in* ‘Proc. Interspeech’, Vol. 15, Singapore, pp. 1964–1968.

Fant, G. (1995), ‘The lf-model revisited. transformations and frequency domain analysis’, *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm* **2**, 3.

Fant, G. & Lin, Q. (1987), ‘Glottal source - vocal tract acoustic interaction’, *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm* **28**(1), 013–027.

Gales, M. J. F. (1999), ‘Semi-tied covariance matrices for hidden Markov models’, *IEEE Trans. Speech Audio Process.* **7**(3), 272–281.

Hanson, H. M. (1997), ‘Glottal characteristics of female speakers: Acoustic correlates’, *J. Acoust. Soc. Am.* **101**(1), 466–481.

Hashimoto, K., Oura, K., Nankaku, Y. & Tokuda, K. (2015), The effect of neural networks in statistical parametric speech synthesis, *in* ‘Proc. ICASSP’, Vol. 40, Brisbane, Australia, pp. 4455–4459.

Henter, G. E., Merritt, T., Shannon, M., Mayo, C. & King, S. (2014), Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech, *in* ‘Proc. Interspeech’, pp. 1504–1508.

Henter, G. E., Ronanki, S., Watts, O., Wester, M., Wu, Z. & King, S. (2016), Robust TTS duration modelling using DNNs, *in* ‘Proc. ICASSP’, Vol. 41, Shanghai, China, pp. 5130–5134.

Hinterleitner, F. (2017), Quality of Synthetic Speech: Perceptual Dimensions, Influencing Factors, and Instrumental Assessment, PhD thesis, TU Berlin.

Hochreiter, S. & Schmidhuber, J. (1997), ‘Long short-term memory’, *Neural Comput.* **9**(8), 1735–1780.

Holm, S. (1979), ‘A simple sequentially rejective multiple test procedure’, *Scand. J. Stat.* **6**(2), 65–70.

Hu, Q., Richmond, K., Yamagishi, J. & Latorre, J. (2013), An experimental comparison of multiple vocoder types, *in* ‘Proc. ISCA SSW’, Vol. 8, Barcelona, Spain, pp. 155–160.

Hu, Q., Yamagishi, J., Richmond, K., Subramanian, K. & Stylianou, Y. (2016), Initial investigation of speech synthesis based on complex-valued neural networks, *in* ‘Proc. ICASSP’, pp. 5630–5634.

Ishi, C. T., Ishiguro, H. & Hagita, N. (2010), ‘Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech’, *EURASIP J. Audio Spee.* **2010**(1).

ITU-R (2015), ‘Method for the subjective assessment of intermediate quality level of audio systems’, ITU Recommendation ITU-R BS.1534-3.

ITU-T (1996), ‘Methods for subjective determination of transmission quality’, ITU Recommendation ITU-T P.800.

Jiang, Y., Ling, Z.-H., Lei, M., Wang, C.-C., Heng, L., Hu, Y., Dai, L.-R. & Wang, R.-H. (2010), The USTC system for Blizzard Challenge 2010, *in* ‘Proc. Blizzard Challenge Workshop’, Vol. 6, Kansai Science City, Japan.

Juvela, L., Bollepalli, B., Airaksinen, M. & Alku, P. (2016), High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network, *in* ‘Proc. ICASSP’, Vol. 41, Shanghai, China, pp. 5120–5124.

Kane, J., Drugman, T. & Gobl, C. (2013), ‘Improved automatic detection of creak’, *Comput. Speech Lang.* **27**(4), 1028–1047.

Karhila, R., Remes, U. & Kurimo, M. (2014), ‘Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments’, *IEEE J. Sel. Topics Signal Process.* **8**(2), 285–295.

Kawahara, H. (2006), ‘STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds’, *Acoust. Sci. Technol.* **27**(6), 349–353.

King, S. (2014), ‘Measuring a decade of progress in text-to-speech’, *Loquens* **1**(1).

Klatt, D. H. & Klatt, L. C. (1990), ‘Analysis, synthesis, and perception of voice quality variations among female and male talkers’, *J. Acoust. Soc. Am.* **87**(2), 820–857.

Kominek, J., Schultz, T. & Black, A. W. (2008), Synthesizer voice quality on new languages calibrated with mean mel-cepstral distortion, *in* ‘Proc. SLTU’, Vol. 1, Hanoi, Viet Nam, pp. 63–68.

Koriyama, T., Nose, T. & Kobayashi, T. (2014), ‘Statistical parametric speech synthesis based on Gaussian process regression’, *IEEE J. Sel. Topics Signal Process.* **8**(2), 173–183.

Latorre, J., Gales, M. J. F., Buchholz, S., Knill, K., Tamurd, M., Ohtani, Y. & Akamine, M. (2011), Continuous f0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?, *in* ‘Proc. ICASSP’, pp. 4724–4727.

Laver, J. (2009), *The Phonetic Description of Voice Quality*, Cambridge University Press.

Ling, Z.-H. & Dai, L.-R. (2012), ‘Minimum Kullback-Leibler divergence parameter generation for HMM-based speech synthesis’, *IEEE Audio, Speech, Language Process.* **20**(5), 1492–1502.

Ling, Z.-H., Deng, L. & Yu, D. (2013), ‘Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis’, *IEEE Audio, Speech, Language Process.* **21**(10), 2129–2139.

Ling, Z.-H., Lu, H., Hu, G.-P., Dai, L.-R. & Wang, R.-H. (2008), The USTC system for Blizzard Challenge 2008, *in* ‘Proc. Blizzard Challenge Workshop’, Vol. 4, Brisbane, Australia.

Ling, Z.-H., Qin, L., Lu, H., Gao, Y., Dai, L.-R., Wang, R.-H., Jiang, Y., Zhao, Z.-W., Yang, J.-H., Chen, J. & Hu, G.-P. (2007), The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007, *in* ‘Proc. Blizzard Challenge Workshop’, Vol. 3, Bonn, Germany.

Lu, H., Ling, Z.-H., Lei, M., Wang, C.-C., Zhao, H.-H., Chen, L.-H., Hu, Y., Dai, L.-R. & Wang, R.-H. (2009), The USTC system for Blizzard Challenge 2009, *in* ‘Proc. Blizzard Challenge Workshop’, Vol. 5, Edinburgh, UK.

McAulay, R. & Quatieri, T. (1986), ‘Speech analysis/synthesis based on a sinusoidal representation’, *IEEE Trans. Acoust., Speech, Signal Process.* **34**(4), 744–754.

Merritt, T. (2016), Overcoming the limitations of statistical parametric speech synthesis, PhD thesis, School of Informatics, The University of Edinburgh, Edinburgh, UK.

Merritt, T., Clark, R. A. J., Wu, Z., Yamagishi, J. & King, S. (2016), Deep neural network-guided unit selection synthesis, *in* ‘Proc. ICASSP’, Vol. 41, Shanghai, China, pp. 5145–5149.

Merritt, T. & King, S. (2013), Investigating the shortcomings of HMM synthesis, *in* ‘Proc. ISCA SSW’, Vol. 8, Barcelona, Spain, pp. 185–190.

Merritt, T., Latorre, J. & King, S. (2015), Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech, *in* ‘Proc. ICASSP’, Vol. 40, Brisbane, Australia, pp. 4220–4224.

Merritt, T., Raitio, T. & King, S. (2014), Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis, *in* ‘Proc. Interspeech’, Singapore, pp. 1509–1513.

Merritt, T., Yamagishi, J., Wu, Z., Watts, O. & King, S. (2015), Deep neural network context embeddings for model selection in rich-context HMM synthesis, *in* ‘Proc. Interspeech’, Dresden, Germany, pp. 1509–1513.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), ‘Efficient estimation of word representations in vector space’, *arXiv preprint arXiv:1301.3781* .

Moore, R. K. (2012), 'A Bayesian explanation of the 'uncanny valley' effect and related psychological phenomena', *Sci. Report.* **2**(864).

Mori, M. (1970), 'Bukimi no tani (the uncanny valley)', *Energy* **7**, 33–35.

Nose, T. (2016), 'Efficient implementation of global variance compensation for parametric speech synthesis', *IEEE/ACM Audio, Speech, Language Process.* **24**(10), 1694–1704.

Pollet, V. & Breen, A. P. (2008), Synthesis by generation and concatenation of multiform segments, *in* 'Proc. Interspeech', pp. 1825–1828.

Qian, Y., Yin, M., You, Y. & Yu, K. (2015), Multi-task joint-learning of deep neural networks for robust speech recognition, *in* 'Proc. IEEE ASRU Workshop', IEEE, pp. 310–316.

Raitio, T., Lu, H., Kane, J., Suni, A., Vainio, M., King, S. & Alku, P. (2014), Voice source modelling using deep neural networks for statistical parametric speech synthesis, *in* 'Proc. EUSIPCO', Vol. 22, Lisbon, Portugal, pp. 2290–2294.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. & Alku, P. (2011), 'HMM-based speech synthesis utilizing glottal inverse filtering', *IEEE Trans. Audio, Speech, Language Process.* **19**(1), 153–165.

Ramamoorthy, V. & Jayant, N. S. (1984), 'Enhancement of ADPCM speech by adaptive postfiltering', *AT&T Tech. J.* **63**(8), 1465–1475.

Rendel, A., Fernandez, R., Hoory, R. & Ramabhadran, B. (2016), Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end, *in* 'Proc. ICASSP', Vol. 41, Shanghai, China, pp. 5655–5659.

Ribeiro, M. S., Yamagishi, J. & Clark, R. A. J. (2015), A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis, *in* 'Proc. Interspeech', Vol. 16, Dresden, Germany, pp. 1586–1590.

Ronanki, S., Watts, O., King, S. & Henter, G. E. (2016), Median-based generation of synthetic speech durations using a non-parametric approach, *in* 'Proc. SLT', San Diego, CA.

Shannon, M. (2014), Probabilistic acoustic modelling for parametric speech synthesis, PhD thesis, Department of Engineering, University of Cambridge, Cambridge, UK.

Shannon, M. & Byrne, W. (2013), Fast, low-artifact speech synthesis considering global variance, *in* 'Proc. ICASSP', Vol. 38, Vancouver, BC, pp. 7869–7873.

Shannon, M., Zen, H. & Byrne, W. (2011), The effect of using normalized models in statistical speech synthesis, *in* 'Proc. Interspeech', Vol. 10, Florence, Italy, pp. 121–124.

Silén, H., Helander, E., Nurminen, J. & Gabbouj, M. (2012), Ways to implement global variance in statistical speech synthesis, *in* 'Proc. Interspeech', Vol. 13, Portland, OR, pp. 1436–1439.

Stylianou, Y. (1996), Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification, PhD thesis, TelecomParis, France.

Takaki, S., Kim, S., Yamagishi, J. & Kim, J. (2015), Multiple feed-forward deep neural networks for statistical parametric speech synthesis, *in* 'Proc. Interspeech', Vol. 16, Dresden, Germany, pp. 2242–2246.

Takashi, N., Tachibana, M. & Kobayashi, T. (2009), 'HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation', *IEICE Trans. Inf. Syst.* **92**(3), 489–497.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. (2011), 'How to grow a mind: Statistics, structure, and abstraction', *Science* **331**(6022), 1279–1285.

Toda, T., Muramatsu, T. & Banno, H. (2012), Implementation of Computationally Efficient Real-Time Voice Conversion, *in* 'Proc. Interspeech', Vol. 13, Portland, OR, pp. 94–97.

Toda, T. & Tokuda, K. (2007), 'A speech parameter generation algorithm considering global variance for HMM-based speech synthesis', *IEICE Trans. Inf. Syst.* **90**(5), 816–824.

Tokuda, K., Hashimoto, K., Oura, K. & Nankaku, Y. (2016), Temporal modeling in neural network based statistical parametric speech synthesis, *in* 'Proc. ISCA SSW', Vol. 9, Sunnyvale, CA, pp. 113–118.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. (2000), Speech parameter generation algorithms for HMM-based speech synthesis, *in* 'Proc. ICASSP', Vol. 25, Istanbul, Turkey, pp. 1315–1318.

Tokuda, K. & Zen, H. (2015), Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis, *in* 'Proc. ICASSP', Vol. 40, Brisbane, Australia, pp. 4215–4219.

Tokuda, K. & Zen, H. (2016), Directly modeling voiced and unvoiced components in speech waveforms by neural networks, *in* 'Proc. ICASSP', Vol. 41, Shanghai, China, pp. 5640–5644.

Uría, B., Murray, I. & Larochelle, H. (2014), A deep and tractable density estimator, *in* 'Proc. ICML', Vol. 31, Beijing, China, pp. 467–475.

Uría, B., Murray, I., Renals, S., Valentini-Botinhao, C. & Bridle, J. (2015), Modelling acoustic feature dependencies with artificial neural networks: Trajectory-RNADE, *in* 'Proc. ICASSP', Vol. 40, Brisbane, Australia, pp. 4465–4469.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016), 'WaveNet: A generative model for raw audio', *arXiv preprint arXiv:1609.03499* .

Wang, W., Xu, S. & Xu, B. (2016), Gating recurrent mixture density networks for acoustic modeling in statistical parametric speech synthesis, *in* 'Proc. ICASSP', Vol. 41, Shanghai, China, pp. 5520–5524.

Wang, X., Takaki, S. & Yamagishi, J. (2016*a*), A comparative study of the performance of HMM, DNN, and RNN based speech synthesis systems trained on very large speaker-dependent corpora, *in* 'Proc. ISCA SSW', Vol. 9, Sunnyvale, CA, pp. 125–128.

Wang, X., Takaki, S. & Yamagishi, J. (2016*b*), Enhance the word vector with prosodic information for the recurrent neural network based TTS system, *in* 'Proc. Interspeech', Vol. 17, San Francisco, CA, pp. 2856–2860.

Watts, O., Henter, G. E., Merritt, T., Wu, Z. & King, S. (2016), From HMMs to DNNs: where do the improvements come from?, *in* 'Proc. ICASSP', Vol. 41, Shanghai, China, pp. 5505–5509.

Watts, O., Yamagishi, J. & King, S. (2010), Letter-based speech synthesis, *in* 'Proc. Speech Synthesis Workshop 2010', Nara, Japan, pp. 317–322.

Wester, M., Valentini-Botinhao, C. & Henter, G. E. (2015), Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations, *in* 'Proc. Interspeech', Vol. 16, Dresden, Germany, pp. 3476–3480.

Wester, M., Watts, O. & Henter, G. E. (2016), Evaluating comprehension of natural and synthetic conversational speech, *in* 'Proc. Speech Prosody', Vol. 8, Boston, MA, pp. 736–740.

Wu, Z. & King, S. (2016), 'Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training', *IEEE/ACM Audio, Speech, Language Process.* **24**(7), 1255–1265.

Wu, Z., Valentini-Botinhao, C., Watts, O. & King, S. (2015), Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis, *in* 'Proc. ICASSP', Vol. 40, Brisbane, Australia, pp. 4460–4464.

Wu, Z., Watts, O. & King, S. (2016), Merlin: An open source neural network speech synthesis system, *in* 'Proc. ISCA SSW', Vol. 9, Sunnyvale, CA, pp. 218–223.

Yamagishi, J., Ling, Z.-H. & King, S. (2008), Robustness of HMM-based speech synthesis, *in* 'Proc. Interspeech', Vol. 9, Brisbane, Australia, pp. 581–584.

Yan, Z.-J., Qian, Y. & Soong, F. K. (2009), Rich context modeling for high quality HMM-based TTS, *in* ‘Proc. Interspeech’, Vol. 10, Brighton, UK, pp. 1755–1758.

Yoshimura, T., Henter, G. E., Watts, O., Wester, M., Yamagishi, J. & Tokuda, K. (2016), A hierarchical predictor of synthetic speech naturalness using neural networks, *in* ‘Proc. Interspeech’, Vol. 17, San Francisco, CA, pp. 342–346.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (2005), ‘Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis’, *Syst. Comput. Jpn.* **36**(12), 43–50.

Yu, K. & Young, S. (2011), ‘Continuous f0 modeling for HMM-based statistical parametric speech synthesis’, *IEEE Trans. Audio, Speech, Language Process.* **19**(5), 1071–1079.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. & Tokuda, K. (2007), The HMM-based speech synthesis system (HTS) version 2.0, *in* ‘Proc. ISCA SSW’, Vol. 6, Bonn, Germany, pp. 294–299.

Zen, H. & Sak, H. (2015), Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis, *in* ‘Proc. ICASSP’, Vol. 40, Brisbane, Australia, pp. 4470–4474.

Zen, H. & Senior, A. (2014), Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis, *in* ‘Proc. ICASSP’, Vol. 39, Florence, Italy, pp. 3844–3848.

Zen, H., Senior, A. & Schuster, M. (2013), Statistical parametric speech synthesis using deep neural networks, *in* ‘Proc. ICASSP’, Vol. 38, Vancouver, BC, pp. 7962–7966.

Zen, H., Tokuda, K. & Black, A. W. (2009), ‘Statistical Parametric Speech Synthesis’, *Speech Comm.* **51**(11), 1039–1064.

Zen, H., Tokuda, K. & Kitamura, T. (2007), ‘Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences’, *Comput. Speech Lang.* **21**(1), 153–173.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (2007), ‘A hidden semi-Markov model-based speech synthesis system’, *IEICE Trans. Inf. Syst.* **90**(5), 825–834.

Zhang, M., Tao, J., Jia, H. & Wang, X. (2008), Improving HMM based speech synthesis by reducing over-smoothing problems, *in* ‘Proc. ISCSLP’, Vol. 6, Kunming, China, pp. 1–4.