Measuring a decade of progress in Text-to-Speech

Measuring a decade of progress in Text-to-Speech

Simon King

The Centre for Speech Technology Research, The University of Edinburgh, UK

Simon.King@ed.ac.uk

ABSTRACT

The Blizzard Challenge offers a unique insight into progress in text-to-speech synthesis over the last decade. By using a very large listening test to compare the performance of a wide range of systems that have been constructed using a common corpus of speech recordings, it is possible to make some direct comparisons between competing techniques. By reviewing over a hundred papers describing all entries to the Challenge since 2005, we can make a useful summary of the most successful techniques adopted by participating teams, as well as drawing some conclusions about where the Blizzard Challenge has succeeded, and where there are still open problems in cross-system comparisons of text-to-speech synthesisers.

Keywords: Text-to-Speech Synthesis, Evaluation, The Blizzard Challenge

1. INTRODUCTION

The last ten years have seen considerable improvements in the quality of speech generated by text-to-speech (TTS) systems, and we have evidence for this from the Blizzard Challenge[1] and the associated summary papers by the organisers (Black and Tokuda, 2005a; Bennett, 2005; Bennett and Black, 2006; Fraser and King, 2007; Karaiskos et al., 2008; King and Karaiskos, 2009, 2010, 2011, 2012, 2013; Prahallad et al., 2013).

1.1. The Blizzard Challenge

Inspired by corresponding evaluation methods in automatic speech recognition (ASR), the Blizzard Challenge (or "Blizzard" in short) set out to provide direct comparisons between systems in a way that was not possible before. As we will briefly describe in Section 1.2, TTS systems are generally rather complex and even messy (to the point of being impossible to optimise in any formal sense) because they rely on a large and disparate collection of linguistic resources and data in order to achieve the difficult transformation from written to spoken language. Blizzard performs cross-system comparisons, and tries to make them as meaningful as possible.

Blizzard is an annual event, started in 2005, in which typically 10 to 20 groups independently build synthetic voices from a common speech corpus and then submit synthetic speech samples to a common evaluation, which uses a large pool of listeners. We will summarise the methodology used by Blizzard in Section 2 and at the end of the paper in Section 4 we will provide a critique of this methodology's strengths and weaknesses. In between, the core of this paper in Section 3 lists the key findings from nearly a decade of Blizzard Challenges: this means identifying the techniques used by the most successful systems which went on to be widely adopted.

This is certainly not a survey of the entire field of speech synthesis – for that you might turn to (Taylor, 2009) for a comprehensive textbook or to (Suendermann et al., 2006) for a discussion of open challenges. Rather, this is a view taken through the lens of the Blizzard Challenge, the only place where direct comparisons across a wide range of systems can be seen.

1.2. The typical architecture of a Text-to-Speech system

In order to understand the scope of the Blizzard Challenge, and in particular what it is able to evaluate and what it so far has not attempted to evaluate, we need to describe a typical TTS system architecture. Almost invariably, systems

[1] http://www.synsig.org/index.php/Blizzard_Challenge

are divided into two components. The first is a linguistic processor, or "front end" which takes unnormalised text and produces from it a "linguistic specification". This will contain information such as a phonetic string, syllabification of that string, some representation of prosody (e.g., accents and boundaries), and so on. The second component is a waveform generator that takes as input this linguistic specification and creates a corresponding speech waveform.

The methods used within the front end are many and various, including both human-created resources such as text normalisation rules or pronunciation dictionaries, and learned-from-data models such as those needed to predict the pronunciation of words not in the dictionary. There are really only two good methods available for the waveform generator: either fragments of recorded speech are selected from a database and concatenated – the unit selection approach – or a (statistical) model learned from that database is used to generate synthetic waveforms via a vocoder. The database is a critical component, and great care is usually taken both in selecting what should be contained in it, and then in recording a professional speaker under ideal studio conditions.

1.3. What this means for any attempt at cross-system evaluation

It is now clear that making comparisons between different TTS systems is not going to be easy, because their performance rests on so many sub-components, any of which could be responsible for differences in the generated speech. In particular, if two systems employ recordings of different speakers then all comparisons may be rendered meaningless because listeners may simply prefer one speaker over the other. It is this factor that Blizzard first set out to control, by using the same speaker in all systems to be compared. Blizzard also controls the database content, by distributing a single shared corpus of speech recordings, often provided from an established company or research group; for example, a corpus was released by ATR for the 2007 Challenge (Ni et al., 2007).

2. THE BLIZZARD CHALLENGE METHODOLOGY

Given the complicated nature of the front end, and the fact that the content of the linguistic specification varies from one system to another, it is hard to design an evaluation that targets the front end specifically. Likewise, since the waveform generation component may be carefully tuned to use one particular form of linguistic specification (particularly in the case of unit selection systems), it is hard to evaluate that in isolation too. So, the Blizzard Challenge is obliged to take a holistic approach and it generally evaluates entire end-to-end systems.

2.1. Common data

The methodology used in the Challenge is described by Black and Tokuda (2005b) and we will only summarise it briefly here. First, a language (or in some years multiple languages) are selected and common data sets are defined. The data minimally comprise recorded speech from a single speaker alongside text transcriptions. Optionally, alignments between the text and speech are provided, possibly including phonetic segmentation or other linguistic annotations on the text such as syllabification, up to and including a complete linguistic specification. Rules on the use of the data, and what additional resources may or may not be employed by participants are defined and refined each year.

2.2. Open participation

An open invitation for participation is sent to the speech synthesis research community, and teams register. During a defined time period, usually of a few months, each team builds their system using the common data. At the end of this period, a set of previously-unseen test material is circulated and teams return the corresponding synthetic speech from their systems.

2.3. Evaluation using a listening test

The organisers conduct a large scale listening test, typically with over 500 listeners, and provide the results to the teams. The Challenge concludes with a workshop and published pa-

pers summarising these results.

2.4. Anonymity

In order to encourage industry participation, and system of anonymity is adopted so that, although the names of all participating teams are made public at the end of the Challenge, the results are presented without showing the correspondence between team names and results in any publication. Individual teams of course know which results are for their system; they may choose to reveal this in their own publications, but this is not required.

3. TECHNIQUES EMPLOYED BY PARTICIPATING SYSTEMS

We now proceed to the main point of this paper: a kind of 'executive summary' of the techniques used by participating teams that have proved most successful and have therefore been widely adopted. The Blizzard Challenge cannot claim to have caused the emergence of new techniques: its claim is more limited and concerns providing independent evidence about the relative merits of competing techniques. This evidence is sometimes more compelling than that found in individual papers because of the direct comparisons made between 'best in class' systems, and the comparisons with natural speech, rather than the usual comparisons made between a single proposed method and a baseline system which is usually also created by the same researchers. The best example of this kind of evidence is the landmark finding that a statistical parametric synthesiser was as intelligible as natural speech and more intelligible than all unit selection systems.

Whilst the first two techniques listed in the next part of the paper (Section 3.1) emerged well before the start of the Blizzard Challenge, they have continued to perform well and can each claim to be "better" than the other along some dimension of the evaluation. Indeed, another good example of the strong evidence that the Challenge provides concerns the relative naturalness and intelligibility of unit selection and statistical parametric approaches.

Applications of TTS Most TTS systems aim at some non-existent 'general purpose' application, but the Blizzard Challenge has also witnessed more targeted systems, such as personalised synthesis for clinical applications (Bunnell et al., 2005, 2010) – something that Yamagishi's adaptive systems (Yamagishi et al., 2007, 2008) are also being used for. Recent Challenges have used audiobooks as a source of transcribed speech recordings and part of the evaluation has involved synthesis of paragraph-sized texts, roughly approximating a TTS audiobook application. The Challenge places no constraints on resources other than the few months allowed to build the system and the few days to synthesise the test material: most Blizzard entries are resource-hungry (in terms of memory and/or compute) server-based research systems. There have been only occasional entires that are small footprint / low compute such as that described by Baumgartner et al. (2012), which would be appropriate for embedded applications.

3.1. Waveform generation

3.1.1. Unit selection generates the most natural speech

Consistently, in every challenge, the system that has been rated as the most natural by listeners has always generated the speech signal by concatenating recorded samples of speech. The size of these units has varied somewhat, as have the methods for selecting and concatenating them, but it is striking that listeners consistently rate recorded speech containing inevitable concatenation artefacts as sounding more natural than speech generated using a vocoder. Nevertheless, whilst listeners might say such speech is more natural, they generally find it harder to understand than speech for a vocoder driven by a statistical parametric model.

The Challenge has seen many 'classical' unit selection systems that closely follow Hunt and Black (1996). The most prototypical of these is the Festival system, with its 'multisyn' unit selection engine (Clark et al., 2005, 2006; Richmond et al., 2007). This system was adopted as a benchmark in later challenges, allowing some

limited comparisons across different years of the challenge to be made (e.g., was a system better or worse than Festival?). Another classical unit selection, which like Festival has its roots in earlier ATR systems, is Ximera (Toda et al., 2006).

Many, many other similar unit selection systems have been entered into the Challenge, with varying degrees of success. This variation points to the fact that unit selection, whilst appearing to be fairly simple, requires a great deal of engineering skill to obtain really good results. These classical unit selection are those described in Weiss et al. (2007), DRESS (Rosales et al., 2008), ILSP's system (Raptis et al., 2010, 2011, 2012; Chalamandaris et al., 2013), MILE (Kumar et al., 2013), Nokia's NTTS (Ding and Alhonen, 2008), Ogmios (Bonafonte et al., 2007, 2008), RACAI (Boros et al., 2013) SVOX (Wouters, 2007), VoiceText (Jun et al., 2007), and WISTON (Tao et al., 2008, 2009, 2010).

The inevitable creep of statistical techniques As soon as the good performance of HMM-based (Section 3.1.2) and later hybrid (Section 3.1.3) synthesisers was demonstrated, many unit selection systems entered into the Challenge started to adopt statistical methods. Jess was initially a classical unit selection system in its first appearances (Cahill and Carson-Berndsen, 2006, 2007) but later added an HMM-based prosody model (Cahill et al., 2011). OpenMary also evolved from unit selection (Schroeder et al., 2006; Schroeder and Hunecke, 2007) by adding a statistical join model (Schroeder et al., 2008) and continues to participate in the Challenge with both unit selection (Schröder et al., 2009; Charfuelan et al., 2013) and HMM-based (Section 3.1.2) systems. The I^2R system likewise has evolved from classical unit selection (Dong et al., 2008, 2009, 2010) to a system employing HMM-guided unit selection (Dong et al., 2011; Lee et al., 2013). Predating these though, is the clunits system Black and Taylor (1997), first entered in 2008 – see Section 3.1.3.

Newcomers can build great unit selection systems too Many unit selection systems entered into the Challenge do not actually perform any better than Festival, so have to be seen mainly as as learning exercise for the participating teams and not a contribution to knowledge. However, the ability to build excellent unit selection systems *can* be developed independently, as demonstrated by a couple of 'newcomers' (from a speech synthesis community point of view). One notable entry into three of the Blizzard Challenges is the classical unit selection system IVONA (Osowski and Kaszczuk, 2006; Osowski, 2007; Kaszczuk and Osowski, 2009) from a previously little-known Polish company. This system achieved outstanding results; the company was subsequently acquired by Amazon. Another previously little-known company has also entered very respectable unit selection systems into the Challenge: Lessac's method uses a unit called the Lesseme (a kind of phonetic/prosodic-context-dependent unit) to very good effect (Nitisaroj et al., 2010, 2011). The reason that the Lesseme works is probably that it hardcodes some of the key target cost features into the unit type, rather than being radically different from more common units like diphones. What do we learn from such systems? We see that unit selection continues to be the obvious choice if building a commercial product; that, with the right engineers, it delivers very high naturalness. The executive summary is pretty clear: if we don't care about controllability, expressivity, or having a library of many voices, and we have the time, money and the right people to do the engineering, then we should choose unit selection every time.

Taking a little more risk The above systems were entered into the Challenge principally to benchmark them against other systems, although typically participants that take part more than once do generally report that the Challenge has helped them improve their systems. On the other hand, some participants in the Challenge use it as an opportunity to take a little more risk and try new ideas. Cerevoice ex-

perimented with compressed waveforms (Aylett et al., 2007) in one Challenge, and a form of data cleaning based on genre pruning in another (Andersson et al., 2008). Some have even used Blizzard as a way to develop research methodology (Kominek et al., 2005).

Voice conversion Blizzard requires that the entered voice sounds close to the provided speaker, which usually means building a voice on that data. Only two unit selection entries have done differently, by starting from an existing voice. The IBM system of 2005 used speaker transformation (Hamza et al., 2005), and a system based on the Festival front end with the AhoTTS waveform concatenator, which first entered in 2008 (Sainz et al., 2008), also applied voice conversion in 2009 (Sainz et al., 2009).

Non-uniform units For engineering simplicity, most systems employ a single unit type such as the diphone or half-phone, but a few try to extend this to non-uniform units. Examples from the Blizzard Challenge include Ding and Alhonen (2007), Yang et al. (2006) which also employs an HMM-generated prosody target, the DSSP system (Latacz et al., 2008) which later added a statistical target cost (Latacz et al., 2009) and trainable context-dependent target cost weights (Latacz et al., 2010), and a system using syllable-sized units plus back-off (Raghavendra et al., 2008).

Learning the unit type and constructing synthetic units Almost all unit selection systems used expert-defined types (e.g., diphones) as the acoustic unit. Two exceptions to this are the IBM unit selection systems which use HMM state-sized units (a fraction of a phone) and employ HMM state clustering to identify classes of interchangeable units (Eide et al., 2006; Fernandez et al., 2008).

Another departure from the usual type of unit is Toshiba's 'plural unit selection and fusion' approach which constructs units by automatically merging together several recorded instances (Buchholz et al., 2007; Li et al., 2008, 2009). Other systems also try to overcome the limitations of units available in the original recordings by constructing additional units either through concatenation (Aylett et al., 2006) or using HMMs (Aylett and Pidcock, 2009), in an offline procedure known as 'bulking'. It's worth re-iterating at this point that we are only concerned in this paper with systems entered into the Blizzard Challenge, and are not attempting to trace ideas back to their inventors.

3.1.2. Statistical parametric methods generate the most intelligible speech

In contrast to the unit selection approach, systems which employ statistical parametric models to drive a vocoder are generally rated as less natural-sounding by listeners. Nevertheless, the same listeners can transcribe this 'less natural' speech more accurately than unit selection output. The Blizzard Challenge has witnessed the most important period of progress for statistical parametric models. The first Challenge already saw the use of the high-quality vocoder that has become the most widely used (STRAIGHT) and explicit duration models (hidden semi-Markov models: HSMMs) (Zen and Toda, 2005) and subsequent years saw systems employing a vast array of enhancements such as MGC-LSP acoustic features which combine the benefits of cepstral and all-pole representations of the spectral envelope, and global variance (GV) (Zen et al., 2006), minimum generation error training (MGE) (Ling et al., 2006, 2007), formant enhancement (Oura et al., 2009), trajectory training (Maia et al., 2009), the use of GV during training along with trainable mixed excitation (Shiga et al., 2010), minimum generation error linear regression (MGELR) model adaptation (Oura et al., 2010), adjustments to the perceptual scales used to represent acoustic features (Yamagishi and Watts, 2010), deterministic annealing expectation maximisation (Hashimoto et al., 2011) and 'chapter-adaptive training' to cope with changes in recording conditions within audiobook training data (Takaki et al., 2013).

Adaptive models High intelligibility might be a very attractive property, but was discovered in the course of evaluation and was not specifically designed or claimed as a feature of these systems. On the other hand, a 'killer feature' of the statistical parametric framework, that is designed right into the system and is one of the main claims of proponents of the statistical approach, is the ability to modify the underlying model parameters. This is most commonly achieved using adaptation techniques borrowed from ASR, then subsequently extended for TTS. Blizzard entries have used supervised speaker adaptation (Yamagishi et al., 2007, 2008) as well as unsupervised adaptation (i.e., with word transcripts obtained using ASR) (Yamagishi et al., 2009), as an effective way to leverage pre-existing recordings of other speakers when constructing a voice for that year's target speaker.

The spread of statistical parametric synthesis Because almost all of the incremental advancements in statistical parametric modelling techniques are implemented in the HTS toolkit, they are available to everyone. This has spawned a great number of entries from what we might call 'HTS users' – groups that use the toolkit in an essentially unmodified form. Entries to the Challenge in this category include Scholtz et al. (2008), Liao and Wu (2009); Liao et al. (2010, 2012); Liao and Pan (2013), Louw et al. (2010, 2013), Nokia's system combining their front end with HTS (Zhang et al., 2010), Cotescu (2011), some recent MARY entries (Charfuelan, 2012; Charfuelan et al., 2013), one of the entries from I^2R (Lee et al., 2013) and a system which combined the flite front end with HTS (Dinh et al., 2013).

As with unit selection, not all of these are better than the HTS benchmark (employed alongside the Festival benchmark, to give an addition point of calibration from year to year). So, whilst statistical parametric methods might rightly claim to be more 'automatic' than unit selection, nevertheless a high degree of expertise and engineering skill is still required to obtain good results.

Improvements to the vocoder through source modelling The hypothesis that the vocoder is the limiting factor in the naturalness of statistical parametric speech synthesis has led to various attempts to construct improved vocoders. Within the Blizzard Challenge, the most prominent strand of research in this area has focussed on improving the excitation source either by modelling residual signals (Maia et al., 2008, 2009), with a parametric glottal waveform model (Andersson et al., 2009) or by using sampled glottal pulse waveforms as in the GlottHMM system (Suni et al., 2010, 2011, 2012).

3.1.3. Hybrid systems: unit selection guided by a statistical parametric model

In the first few years of the Challenge, it became clear that statistical parametric systems consistently had the better intelligibility, whereas unit selection systems consistently had better naturalness. Although never formally proven, it is widely thought that this better naturalness was a result of using recorded waveforms – in other words, it is a local property of the signal that is partly independent of concatenation artefacts. Conversely, it is widely thought that the intelligibility of statistical parametric systems is a result of their ability to more accurately generate context-dependent speech units (as opposed to the out-of-context units of unit selection). An obvious next step was then to retain unit selection as the method for waveform generation – thus ensuring a natural-sounding signal – but to select the units using a statistical parametric model – thus taking advantage of its ability to predict the acoustic properties of units-in-context that did not occur in the available recorded corpus but that were needed at synthesis time.

Probabilistic models for unit selection The hand-crafted nature of the join and target cost functions used in classical unit selection are often seen as unsatisfactory, since they must be tuned by ear and it is not possible to be sure that optimal values of the various parameters (e.g., weights on linguistic features) have been found. Overcoming this limitation has been a long-

standing goal in unit selection research. Within the Blizzard Challenge, we have observed a number of systems tackling this problem. Sakai and Shu (2005); Sakai (2006) describe a system evolved from MIT's Envoice in which probabilistic models replace almost all hand-tunable parameters. Likewise, the 'clunits' method, first entered to the Challenge in 2008 (Black et al., 2008; Oliveira et al., 2008) builds clustering trees which group together acoustically interchangeable units which share a subset of linguistic feature values. Other attempts at trainable unit selection include the two early entries from μXac (Rozak, 2007, 2008) followed by the much improved system described in Rozak (2009). Lessac also entered systems in which an acoustic target, in this case from a Hierarchical Mixture of Experts, guided the selection of units (Wilhelms-Tricarico et al., 2012, 2013).

The weakness of most attempts to employ learned-from-data models in unit selection is perhaps that they pay attention only to acoustic similarity and do not involve human perceptual judgments. This is probably why a hand-tuned target cost is still better, if correctly constructed and tuned by an expert: it accounts for perceptual judgements.

Hybrid systems We define 'hybrid' systems as those which employ a statistical parametric model – which is in itself *capable of generating speech* in conjunction with a vocoder – to guide the selection of units from the database, which are subsequently concatenated. There is of course not a clear dividing line: for example, the unit selection system described by Wilhelms-Tricarico et al. (2012, 2013) uses a powerful statistical model to predict an acoustic target trajectory, but without any intention of generating speech from it.

The first proposal of a hybrid system observed in the Blizzard Challenge was from Kominek and Black (2006), who mentioned both 'clunits' and HTS as candidates for the statistical parametric model, but actually used their own 'ClusterGen' method as the statistical parametric component; this is rather similar to decision tree-clustered HMM states, as used in HTS. The system was refined and entered again in 2007 (Black et al., 2007).

Subsequently, the 'hybridistion' of HMM-based synthesis with unit selection was developed and placed on a formal mathematical foundation in which the probabilistic nature of the HMMs was made use of. The sequence of highly-successful entries from USTC and their spinout iFlytek are strong evidence that this technique does indeed combine benefits of unit selection and statistical parametric models (Ling et al., 2007, 2008; Lu et al., 2009; Jiang et al., 2010). Subsequent systems of theirs experimented with Lessemes as the modelling unit (Chen et al., 2011), channel- and expressiveness-related labels for audiobook data (Ling et al., 2012), automatic weight learning based on an objective quality model (Chen et al., 2013) and vocal tract resonance (VTR) trajectory-guided unit selection (Zhang et al., 2009).

In parallel to the USTC/iFlytek system evolution, Microsoft Research Asia (MSRA) have entered similar systems. The rather elegant name of 'trajectory tiling' was coined by them and featured in their 2010 entry to the Challenge (Qian et al., 2010). It alludes to a method used in computer graphics in which a parametric model (e.g., a wireframe or skeleton) is given a 'skin' composed from sampled images. The skeleton is convenient for the artist to manipulate and is flexible enough to produce any desired pose, whilst the detailed skin convinces the viewer that the object is real and not computer-generated. In speech, the corresponding advantages are that the underlying statistical parametric model is able to generate any speech sound in any context (the 'trajectory'), whilst the overlayed samples ('tiles') provide the necessary details to convince the listener that the signal is natural speech.

In latter years, more groups have adopted various forms of the hybrid approach, including the NTNU (Meen and Svendsen, 2010), BUCEADOR (Sainz et al., 2011), and SHRC-Ginkgo systems (Yu et al., 2013).

3.2. Linguistic features

It is impossible, for the reasons discussed in Section 1.2 to make many meaningful comparisons across the linguistic processors employed in the Blizzard Challenge. The differences are numerous and their effects on the speech output are impossible to quantify. This has not prevented us still drawing very concrete conclusions about waveform generation though, because we observe the same patterns in intelligibility and naturalness across multiple systems – employing different front ends – and across several years of the challenge.

All we can do with regard to the linguistic features predicted by each system from the text input is to highlight exceptional or unusual features employed by some systems.

Unsupervised features It should be clear that typical front ends are knowledge-rich and are both difficult and expensive to construct. To sidestep this, the system described by Watts et al. (2013) attempted to predict features from text without requiring any human expertise or pre-built resources such as pronunciation dictionaries. The method failed on English, but was reasonably successful on several more well-behaved languages.

Wider and deeper features With the introduction of audiobook data in the Challenge, the opportunity arose to use information beyond the current sentence, which has been tried in several ways including simply appending them as additional contextual features to HMMs (Takaki et al., 2012). Wider context may also be used to separate out disparate data, such as with the channel- and expressiveness-related labels of Ling et al. (2012), or the 'chapter-adaptive training' to cope with changes in recording conditions within audiobook training data used by Takaki et al. (2013).

Whilst many believe that a 'deeper' analysis of the text should yield useful features, it has proven very hard to obtain measurable improvement in the output speech. A possible exception to this is the excellent system described by Yu et al. (2013), which uses syntactic parser features for an audiobook synthesis task.

4. A CRITIQUE OF THE BLIZZARD CHALLENGE

4.1. Positive contributions

In addition to the unquantifiable warm feeling of improved speech synthesis community cohesion and a spirit of sharing techniques and data, the Blizzard Challenge can claim a couple of concrete contributions in its own right.

4.1.1. Advances in objective measures

Although not directly used to rank the systems with the Challenge, objective measures of speech quality have made some progress over the last decade. Most notable is the work of Falk et al. (2008), Hinterleitner et al. (2010) and Norrenbrock et al. (2012) who have collectively pursued instrumental (that is, signal-based rather than listener-based) measures; these have begun to show useful results. These measures attempt to replicate the judgements that listeners would provide for a given set of speech signals. The Blizzard Challenge has been able to provide a substantial training set of signals-plus-listener-ratings on which object measures can be tuned and additional independent data sets on which their effectiveness can be tested.

4.1.2. Spinoffs and related evaluations

The Blizzard Challenge was itself inspired by the long tradition of common evaluation tasks from the field of ASR, and has in turn inspired others to use this methodology to measure (and hopefully promote) progress in other fields. The Hurricane Challenge (Cooke et al., 2013) evaluated methods for improving the intelligibility of natural or synthetic speech in the presence of additive noise, and its organisation closely followed the Blizzard model, with an open invitation to the community to participate, a common data set and set of rules, and a large centralised listening test run by the organisers. The Albayzin Challenges in 2010 (Díaz et al., 2011) and 2012 included a replication of the Blizzard Challenge, using a Spanish corpus.

4.2. Room for improvement

4.2.1. What to evaluate

Naturalness and intelligibility remain the main evaluation criteria for speech synthesis, with judgements being elicited from listeners on a Lickert scale (Likert, 1932). Naturalness remains poorly defined, although listeners do seem to have a clear idea of what is being asked of them given the consistency of their judgements. Intelligibility is measured, as noted in Section 4.2.2, in a particularly unrealistic, or 'ecologically invalid', way.

Blizzard also adds an evaluation of speaker similarity to the mix. This was introduced initially only as a check that participants were using the provided recordings and not entering pre-built systems. With the advent of speaker-adaptive approaches, and for unit selection entires employing voice conversion, speaker similarity became a useful dimension of the evaluation in its own right.

Despite continued calls by the organisers, few researchers in the community have risen to their challenge to propose new and better listening test designs, and in particular to propose *what* to evaluate. The only exception to this is Hinterleitner et al. (2011), who proposed a multi-dimensional test for evaluating synthetic audiobooks. Their method was adopted by the Blizzard Challenge organisers in those later years where audiobook data was used.

4.2.2. How to evaluate

Playing synthetic speech to listeners and asking them to make some response (e.g., provide a rating for a specified property) or perform a task (e.g., transcribe the words they heard) is the bread and butter of synthetic speech evaluation. Whilst objective measures have their place in single-system tuning or in identifying gross differences between systems, a listening test remains the only sure way to demonstrate the superiority of one's proposed new method.

The problem of evaluating synthetic speech via listening tests is not a solved one. It is intrinsically difficult for two reasons. First, it is not clear exactly *what* properties to evaluate. Second, it is hard to know *how* to evaluate the chosen properties, and one can never be certain that all of the listeners have correctly performed the task you expected of them.

Blizzard takes a simple approach to alleviating these worries. The instructions given to listeners are generally simple and do not require any training or high level of knowledge on the listeners' part. A large number of listeners is employed, thus minimising the effect of individuals who fail to follow these instructions. The statistical tests for significant differences are deliberately conservative (Clark et al., 2007) in order to avoid false claims. Of course, the flip-side of this is that it is possible Blizzard fails to identify interesting differences some of the time.

The listening tests typically used by the TTS research community lack ecological validity in many ways. They take place in an unusual setting – quiet, comfortable listening booths with high-quality sound reproduction and no distractions – and ask listeners to perform tasks they would never do in everyday life. For example, in order to test the intelligibility of systems, listeners are asked to transcribe – by typing on a computer keyboard – the individual words they heard. It is hard to think of a real application where this would be done. Worse, the sentences played to listeners are deliberately hard to comprehend, often being devoid of meaning (Benoit and Grice, 1996). This is done to remove the ceiling effect: in other words, many synthesisers could be close to 100% intelligible if predictable, meaningful sentences were used.

Does the lack of ecological validity matter though? In some respects it certainly is not a problem: if our synthesiser is as intelligible as natural speech when using difficult, meaningless sentences then we would be confident that it would be at least as intelligible using normal sentences. That is, the laboratory testing situation can uncover effects that would shrink into insignificance in the real world and the only danger is that we are identifying rather small differences. We still have confidence that we can identify the best system, although we may over-estimate how much better than the next

system it actually is.

But in other respects the lack of ecological validity is much more serious. The idealised environment is the most serious issue: real end users do not operate in quiet environments free of distractions. The 2009 Challenge included a condition in which the synthetic speech was corrupted by a simulated telephone channel (King and Karaiskos, 2009) and the Hurricane Challenge mentioned in Section 4.1.2 addressed the problem of speech-in-noise much more rigourously. The tasks used are also a problem, since listeners are allowed to perform them under no significant constraints on their attention or time. There is doubtless still much to learn from experimental psychology, including the use of distractors to disguise to true purpose of the experiment, or methods which can introduce realistic levels of cognitive load into our subjects.

Despite these widely-recognised potential problems with how TTS is generally evaluated, there have been few attempts to innovate. Perhaps this is for the simple reason that any alternative would almost certainly yield far fewer data points per hour of testing time than current paradigms, and so be less practical and more costly. But perhaps it is just plain laziness: researchers prefer to spend their time inventing exciting new methods for synthesising speech, not worrying about whether they are actually measuring the quality of their work in the best way, especially when the burden of some of that evaluation can be offloaded to an external Challenge.

4.3. Open issues

4.3.1. Whole system vs. component-level evaluations

As we mentioned in Section 2, Blizzard only attempts end-to-end system evaluations. Moreover, it also bundles in the data preparation stages such as alignment with the text and optional hand-corrections performed by some participants. In other words, it evaluates the totality of the *system components* and the *engineering skill and effort* needed to make it work well on a new database. Conclusions about which *method* is "best" are therefore inevitably filtered through the level of expertise and available resources of the team implementing that method. This may be a partial explanation of the "failure" of some entries: the idea had merit, but the implementation was flawed. The availability of resources for checking and correcting the data varies widely between participants. To quantify the effect this has on overall quality, one year's Challenge did release hand-checked alignments but this was found to be of limited use because it does not guarantee consistency across systems, since some may use a different phonetic inventory or pronunciation dictionary. Some participants have themselves investigated the benefits of manual annotations (Chu et al., 2006).

Providing linguistic specifications may appear to be one way to isolate the waveform generation component, but it would not be possible for some participants to modify their systems to use an externally-provided linguistic specification.

4.3.2. Common data, but what else?

The core of the Blizzard Challenge is the shared corpus which all participants are required to use. Its size has varied over the years, generally getting larger over time, and several years have seen specific sub-challenges involving restricted corpus sizes. As we have mentioned a number of times throughout this paper, a common corpus only 'levels the playing field' to some degree and there remain many other uncontrolled factors which may explain differences between systems. It is probably impossible to entirely separate out the effectiveness of a proposed technique from the skill of the engineer who implements it. Simple techniques, implemented by experts, can perform very well. Certainly, complex techniques poorly implemented are not likely to succeed. Within a single year of the Challenge then, it is hard to say for sure that one technique is better than another.

But, by looking over several years of Challenges, as we have done here, we can start to find independently-constructed systems being entered that use a common technique. When

we see several of these performing well, then it becomes more reasonable to say that this is a good technique. Clear examples of this (if implemented skilfully) include unit selection, which almost guarantees a good naturalness score, HMM-based methods, which almost guarantee good intelligibility, and hybrid systems which maintain the high naturalness of unit selection and start to approach the intelligibility of HMM systems.

4.3.3. Too much at stake leads to too little risk

As the Challenge became more and more established, and a firm fixture in the calendar, awareness of it began to rise outside the immediate circle of participating researchers. A negative effect of this is that participation in the Challenge has become a more public affair: poorly-performing entries no longer go un-noticed but instead start to attract attention. For the research labs in large corporations, this presents a major barrier to participation in the Challenge, since their management/lawyers/marketing department are likely to say "Of course you can enter the Blizzard Challenge, provided that you win."

It is often said that one learns more from mistakes than successes, and Blizzard is no exception. The organisers of Blizzard are always at pains to point out that it is not a competition, and there are no winners and losers – that is, 'mistakes' are encouraged. It is to be hoped that all participants resist the temptation to play it safe with their entries, and that normally risk-averse corporations see the benefits to taking part. They can easily mitigate the risks simply by describing their entry as a highly experimental research idea and not as a production system.

REFERENCES

Andersson, J. S., Badino, L., Watts, O. S., and Aylett, M. P. (2008). The CSTR/Cereproc Blizzard entry 2008: The inconvenient data. In *Blizzard Challenge Workshop 2008*.

Andersson, J. S., Cabral, J. P., Badino, L., Yamagishi, J., and Clark, R. A. J. (2009). Glottal source and prosodic prominence modelling in HMM-based speech synthesis for the Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Aylett, M. P., Andersson, J. S., Badino, L., and Pidcock, C. J. (2007). The Cerevoice Blizzard entry 2007: are small database errors worse than compression artifacts? In *Blizzard Challenge Workshop 2007*.

Aylett, M. P. and Pidcock, C. J. (2009). The CereProc Blizzard entry 2009: Some dumb algorithms that don't work. In *Blizzard Challenge Workshop 2009*.

Aylett, M. P., Pidcock, C. J., and Fraser, M. E. (2006). The Cerevoice Blizzard entry 2006: A prototype database unit selection engine. In *Blizzard Challenge Workshop 2006*.

Baumgartner, M., Wilhelms-Tricarico, R., and Reichenbach, J. (2012). The Lessac Technologies time domain diphone parametric synthesis system for microcontrollers for Blizzard Challenge. In *Blizzard Challenge Workshop 2012*.

Bennett, C. L. (2005). Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Bennett, C. L. and Black, A. W. (2006). Blizzard Challenge 2006: Results. In *Blizzard Challenge Workshop 2006*.

Benoit, C. and Grice, M. (1996). The SUS test: a method for the assessment of text-to-speech intelligibility using semantically unpredictable sentences. *Speech Communication*, 18:381–392.

Black, A. and Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proc. Eurospeech*, volume 2, pages 601–604, Rhodes, Greece.

Black, A. W., Bennett, C. L., Blanchard, B. C., Kominek, J., Langner, B., Prahallad, K., and Toth, A. (2007). CMU Blizzard 2007: a hybrid acoustic unit selection system from statistically predicted parameters. In *Blizzard Challenge Workshop 2007*.

Black, A. W., Bennett, C. L., Kominek, J., Langner, B., Prahallad, K., and Toth, A. (2008). CMU Blizzard 2008: Optimally using a large database for unit selection synthesis. In *Blizzard Challenge Workshop 2008*.

Black, A. W. and Tokuda, K. (2005a). The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Black, A. W. and Tokuda, K. (2005b). The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc Interspeech 2005*, Lisbon.

Bonafonte, A., Adell, J., Agüero, P. D., Erro, D., Esquerra, I., Moreno, A., Peréz, J., and Polyakova, T. (2007). The UPC TTS system description for the 2007 Blizzard Challenge. In *Blizzard Challenge Workshop 2007*.

Bonafonte, A., Moreno, A., Adell, J., Aguero, P. D., Banos, E., Erro, D., Esquerra, I., Perez, J., and Polyakova, T. (2008). The UPC TTS system description for the 2008 Blizzard Challenge. In *Blizzard Challenge Workshop 2008*.

Boros, T., Ion, R., and Dumitrescu, S. D. (2013). The RACAI text-to-speech synthesis system. In *Blizzard Challenge Workshop 2013*.

Buchholz, S., Braunschweiler, N., Morita, M., and Webster, G. (2007). The Toshiba entry for the 2007 Blizzard Challenge. In *Blizzard Challenge Workshop 2007*.

Bunnell, H. T., Pennington, C., Yarrington, D., and Gray, J. (2005). Automatic personal synthetic voice construction. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Bunnell, T., Lilley, J., Pennington, C., Moyers, B., and Polikoff, J. (2010). The ModelTalker system. In *Blizzard Challenge Workshop 2010*.

Cahill, P. and Carson-Berndsen, J. (2006). The Jess Blizzard Challenge 2006 entry. In *Blizzard Challenge Workshop 2006*.

Cahill, P. and Carson-Berndsen, J. (2007). The Jess Blizzard Challenge 2007 entry. In *Blizzard Challenge Workshop 2007*.

Cahill, P., Ogbureke, U., Cabral, J., Szekely, E., Abou-Zleikha, M., Ahmed, Z., and Carson-Berndsen, J. (2011). UCD Blizzard Challenge 2011 entry. In *Blizzard Challenge Workshop 2011*.

Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., and Raptis, S. (2013). The ILSP/INNOETICS text-to-speech system for the Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Charfuelan, M. (2012). MARY TTS HMM-based voices for the Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

Charfuelan, M., Pammi, S., and Steiner, I. (2013). MARY TTS unit selection and HMM-based voices for the Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Chen, L.-H., Ling, Z.-H., Song, Y. J. Y., Xia, X.-J., Zu, Y.-Q., Yan, R.-Q., and Dai, L.-R. (2013). The USTC system for Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Chen, L.-H., Yang, C.-Y., Ling, Z.-H., Jiang, Y., Dai, L.-R., Hu, Y., and Wang, R.-H. (2011). The USTC system for Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Chu, M., Chen, Y., Zhao, Y., Li, Y., and Soong, F. (2006). A study on how human annotations benefit the TTS voice. In *Blizzard Challenge Workshop 2006*.

Clark, R., Richmond, K., Strom, V., and King, S. (2006). Multisyn voice for the Blizzard Challenge 2006. In *Blizzard Challenge Workshop 2006*.

Clark, R. A., Richmond, K., and King, S. (2005). Multisyn voices from ARCTIC data for the Blizzard Challenge. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Clark, R. A. J., Podsiadlo, M., Fraser, M., Mayo, C., and King, S. (2007). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Blizzard Challenge Workshop 2007*.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (2013). Intelligibility-enhancing speech modifications: the Hurricane Challenge. In *Proc. Interspeech*, Lyon, France.

Cotescu, M. (2011). PUB entry in the Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Díaz, F. C., Pazó, F. J. M., Arza, M., Fernández, L. D., Bonafonte, A., Navas, E., and Sainz, I. (2011). Albayzín 2010: A Spanish text to speech evaluation. In *Proc Interspeech 2011*, Florence, Italy.

Ding, F. and Alhonen, J. (2007). Non-uniform unit selection through search strategy for Blizzard Challenge 2007. In *Blizzard Challenge Workshop 2007*.

Ding, F. and Alhonen, J. (2008). NTTS participation in the Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Dinh, A.-T., Phan, T.-S., Phan, D.-H., Phi, T.-L., Vu, T.-T., and Luong, C.-M. (2013). The iSolar Blizzard Challenge 2013 entry. In *Blizzard Challenge Workshop 2013*.

Dong, M., Cen, L., Chan, P., Huang, D., Zhu, D., Ma, B., and Li, H. (2009). I^2R text-to-speech system for Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Dong, M., Chan, P., Cen, L., Ma, B., and Li, H. (2010). I^2R text-to-speech system for Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Dong, M., Lee, S. W., Chan, P., and Cen, L. (2011). I^2R text-to-speech system for Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Dong, M., Zhu, D., Ma, B., and Li, H. (2008). I^2R's submission to Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Eide, E., Fernandez, R., Hoory, R., Hamza, W., Kons, Z., Picheny, M., Sagi, A., Shechtman, S., and Shuang, Z. W. (2006). The IBM submission to the 2006 Blizzard text-to-speech Challenge. In *Blizzard Challenge Workshop 2006*.

Falk, T. H., Moeller, S., Karaiskos, V., and King, S. (2008). Improving instrumental quality prediction performance for the Blizzard Challenge. In *Blizzard Challenge Workshop 2008*.

Fernandez, R., Kons, Z., Shechtman, S., Shuang, Z. W., Hoory, R., Ramabhadran, B., and Qin, Y. (2008). The IBM submission to the 2008 text-to-speech Blizzard Challenge. In *Blizzard Challenge Workshop 2008*.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Blizzard Challenge Workshop 2007*.

Hamza, W., Bakis, R., Shuang, Z. W., and Zen, H. (2005). On building a concatenative speech synthesis system from the Blizzard Challenge speech databases. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Hashimoto, K., Takaki, S., Oura, K., and Tokuda, K. (2011). Overview of NIT HMM-based speech synthesis system for Blizzard

Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Hinterleitner, F., Möller, S., Falk, T. H., and Polzehl, T. (2010). Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from Blizzard Challenges 2008 and 2009. In *Blizzard Challenge Workshop 2010*.

Hinterleitner, F., Neitzel, G., Möller, S., and Norrenbrock, C. (2011). An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Blizzard Challenge Workshop 2011*.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP-96*, pages 373–376, Atlanta, Georgia.

Jiang, Y., Ling, Z.-H., Lei, M., Wang, C.-C., Heng, L., Hu, Y., Dai, L.-R., and Wang, R.-H. (2010). The ustc system for Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Jun, W.-S., Na, D.-S., Kim, S.-W., Kim, M., Lee, J.-W., and Lee, J.-S. (2007). The Voicetext text-to-speech system for the Blizzard Challenge 2007. In *Blizzard Challenge Workshop 2007*.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Kaszczuk, M. and Osowski, L. (2009). The IVO software Blizzard Challenge 2009 entry: Improving IVONA text-to-speech. In *Blizzard Challenge Workshop 2009*.

King, S. and Karaiskos, V. (2009). The Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

King, S. and Karaiskos, V. (2010). The Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

King, S. and Karaiskos, V. (2011). The Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

King, S. and Karaiskos, V. (2012). The Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

King, S. and Karaiskos, V. (2013). The Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Kominek, J., Bennett, C. L., Langner, B., and Toth, A. R. (2005). The Blizzard Challenge 2005 cmu entry – a method for improving speech synthesis systems. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Kominek, J. and Black, A. W. (2006). The Blizzard Challenge 2006 CMU entry introducing hybrid trajectory-selection synthesis. In *Blizzard Challenge Workshop 2006*.

Kumar, H. R. S., Ashwini, J. K., Rajaramand, B. S. R., and Ramakrishnan, A. G. (2013). MILE TTS for tamil and kannada for blizzard challenge 2013. In *Blizzard Challenge Workshop 2013*.

Latacz, L., Kong, Y. O., Mattheyses, W., and Verhelst, W. (2008). An overview of the VUB entry for the 2008 Blizzard Challenge. In *Blizzard Challenge Workshop 2008*.

Latacz, L., Mattheyses, W., and Verhelst, W. (2009). The VUB Blizzard Challenge 2009 entry. In *Blizzard Challenge Workshop 2009*.

Latacz, L., Mattheyses, W., and Verhelst, W. (2010). The VUB Blizzard Challenge 2010 entry: Towards automatic voice building. In *Blizzard Challenge Workshop 2010*.

Lee, S. W., Dong, M., Ang, S. T., and Chew, M. M. (2013). I^2R text-to-speech system for Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Li, J., Luan, J., Yi, L., Lou, X., Wang, X., He, L., and Hao, J. (2009). The Toshiba Mandarin

TTS system for the Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Li, J., Xu, D., Yi, L., Lou, X., Luan, J., Wang, X., He, L., and Hao, J. (2008). The Toshiba Mandarin TTS system for the Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Liao, Y.-F., Lin, C.-C., and Pan, J.-Y. (2012). The NTUT Blizzard Challenge 2012 entry. In *Blizzard Challenge Workshop 2012*.

Liao, Y.-F. and Pan, J.-Y. (2013). The NTUT Blizzard Challenge 2013 entry. In *Blizzard Challenge Workshop 2013*.

Liao, Y.-F. and Wu, M.-L. (2009). The NTUT Blizzard Challenge 2009 entry. In *Blizzard Challenge Workshop 2009*.

Liao, Y.-F., Wu, M.-L., and Lyu, S.-H. (2010). The NTUT Blizzard Challenge 2010 entry. In *Blizzard Challenge Workshop 2010*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Ling, Z.-H., Lu, H., Hu, G.-P., and Li-Rong Dai, R.-H. W. (2008). The USTC system for Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

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., and Hu, G.-P. (2007). The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007. In *Blizzard Challenge Workshop 2007*.

Ling, Z.-H., Wu, Y.-J., Wang, Y.-P., Qin, L., and Wang, R.-H. (2006). USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In *Blizzard Challenge Workshop 2006*.

Ling, Z.-H., Xia, X.-J., Song, Y., Yang, C.-Y., Chen, L.-H., and Dai, L.-R. (2012). The USTC system for Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

Louw, J. A., Schlünz, G. I., van der Walt, W., de Wet, F., and Pretorius, L. (2013). The Speect text-to-speech system entry for the Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Louw, J. A., van Niekerk, D. R., and Schlünz, G. I. (2010). Introducing the Speect speech synthesis platform. In *Blizzard Challenge Workshop 2010*.

Lu, H., Ling, Z.-H., Lei, M., Wang, C.-C., Zhao, H.-H., Chen, L.-H., Hu, Y., Dai, L.-R., and Wang, R.-H. (2009). The USTC system for Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Maia, R., Ni, J., Sakai, S., Toda, T., Tokuda, K., Shimizu, T., and Nakamura, S. (2008). The NICT/ATR speech synthesis system for the Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Maia, R., Toda, T., Sakai, S., Shiga, Y., Ni, J., Kawai, H., Tokuda, K., Tsuzaki, M., and Nakamura, S. (2009). The NICT entry for the Blizzard Challenge 2009: an enhanced HMM-based speech synthesis system with trajectory training considering global variance and state-dependent mixed excitation. In *Blizzard Challenge Workshop 2009*.

Meen, D. and Svendsen, T. (2010). The NTNU concatenative speech synthesizer. In *Blizzard Challenge Workshop 2010*.

Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R., and Nakamura, S. (2007). ATRECSS – ATR English speech corpus for speech synthesis. In *Blizzard Challenge Workshop 2007*.

Nitisaroj, R., Wilhelms-Tricarico, R., Mottershead, B., Nitisaroj, R., Baumgartner, M., Reichenbach, J., and Marple, G. (2011). The Lessac Technologies system for Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Nitisaroj, R., Wilhelms-Tricarico, R., Mottershead, B., Reichenbach, J., and Marple, G.

(2010). The Lessac Technologies system for Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Norrenbrock, C. R., Hinterleitner, F., Heute, U., and Möller, S. (2012). Towards perceptual quality modeling of synthesized audiobooks - Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

Oliveira, L. C., Paulo, S., Figueira, L., and Mendes, C. (2008). The INESC-ID Blizzard entry: Unsupervised voice building and synthesis. In *Blizzard Challenge Workshop 2008*.

Osowski, L. and Kaszczuk, M. (2006). IVO Blizzard 2006 entry. In *Blizzard Challenge Workshop 2006*.

Osowski, M. K. L. (2007). The IVO software Blizzard 2007 entry: improving Ivona speech synthesis system. In *Blizzard Challenge Workshop 2007*.

Oura, K., Hashimoto, K., Shiota, S., and Tokuda, K. (2010). Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Oura, K., Wu, Y.-J., and Tokuda, K. (2009). Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Prahallad, K., Vadapalli, A., Elluru, N., Mantena, G., Pulugundla, B., Bhaskararao, P., Murthy, H. A., King, S., Karaiskos, V., and Black, A. W. (2013). The Blizzard Challenge 2013 – Indian language task. In *Blizzard Challenge Workshop 2013*.

Qian, Y., Zhi-Jie Yan, Y.-J. W., Soong, F. K., Zhang, G., and Wang, L. (2010). An HMM trajectory tiling (HTT) approach to high quality TTS - Microsoft entry to Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Raghavendra, E., Desai, S., Yegnanarayana, B., Black, A. W., and Prahallad, K. (2008). Blizzard 2008: Experiments on unit size for unit selection speech synthesis. In *Blizzard Challenge Workshop 2008*.

Raptis, S., Chalamandaris, A., Tsiakoulis, P., and Karabetsos, S. (2010). The ILSP text-to-speech system for the Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Raptis, S., Chalamandaris, A., Tsiakoulis, P., and Karabetsos, S. (2011). The ILSP text-to-speech system for the Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Raptis, S., Chalamandaris, A., Tsiakoulis, P., and Karabetsos, S. (2012). The ILSP text-to-speech system for the Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

Richmond, K., Strom, V., Clark, R. A., Yamagishi, J., and Fitt, S. (2007). Festival Multisyn voices for the 2007 Blizzard Challenge. In *Blizzard Challenge Workshop 2007*.

Rosales, H. G., Jokisch, O., and Hoffmann, R. (2008). The DRESS Blizzard Challenge 2008 entry. In *Blizzard Challenge Workshop 2008*.

Rozak, M. (2007). Text-to-speech designed for a massively multiplayer online role-playing game (MMORPG). In *Blizzard Challenge Workshop 2007*.

Rozak, M. (2008). Circumreality functionality delta: Blizzard Challenge 2007 to 2008. In *Blizzard Challenge Workshop 2008*.

Rozak, M. (2009). CircumReality text-to-speech, a talking speech recognizer. In *Blizzard Challenge Workshop 2009*.

Sainz, I., Erro, D., Navas, E., Adell, J., and Bonafonte, A. (2011). BUCEADOR hybrid TTS for Blizzard Challenge 2011. In *Blizzard Challenge Workshop 2011*.

Sainz, I., Erro, D., Navas, E., Hernáez, I., Saratxaga, I., Luengo, I., and Odriozola, I. (2009). The AHOLAB Blizzard Challenge 2009 entry. In *Blizzard Challenge Workshop 2009*.

Sainz, I., Navas, E., and Hernaez, I. (2008). The AHOLAB Blizzard Challenge 2008 entry. In *Blizzard Challenge Workshop 2008*.

Sakai, S. (2006). Building probabilistic corpus-based speech synthesis systems from the Blizzard Challenge 2006 speech databases. In *Blizzard Challenge Workshop 2006*.

Sakai, S. and Shu, H. (2005). A probabilistic approach to unit selection for corpus-based speech synthesis. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Scholtz, P., Visagie, A., and du Preez, J. (2008). Statistical speech synthesis for the Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Schröder, M., Pammi, S., and Türk, O. (2009). Multilingual MARY TTS participation in the Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Schroeder, M., Charfuelan, M., Pammi, S., and Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In *Blizzard Challenge Workshop 2008*.

Schroeder, M. and Hunecke, A. (2007). MARY TTS participation in the Blizzard Challenge 2007. In *Blizzard Challenge Workshop 2007*.

Schroeder, M., Hunecke, A., and Krstulovic, S. (2006). OpenMary - open source unit selection as the basis for research on expressive synthesis. In *Blizzard Challenge Workshop 2006*.

Shiga, Y., Toda, T., Sakai, S., Ni, J., Tokuda, H. K. K., Tsuzaki, M., and Nakamura, S. (2010). NICT Blizzard Challenge 2010 entry. In *Blizzard Challenge Workshop 2010*.

Suendermann, D., Höge, H., and Black, A. (2006). Challenges in speech synthesis. In Chen, F. and Jokinen, K., editors, *Speech Technology*, chapter 2, pages 19–32. Elsevier.

Suni, A., Raitio, T., Vainio, M., and Alku, P. (2010). The glottHMM speech synthesis entry for Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Suni, A., Raitio, T., Vainio, M., and Alku, P. (2011). The glottHMM speech synthesis entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation. In *Blizzard Challenge Workshop 2011*.

Suni, A., Raitio, T., Vainio, M., and Alku, P. (2012). The glottHMM entry for Blizzard Challenge 2012: Hybrid approach. In *Blizzard Challenge Workshop 2012*.

Takaki, S., Sawada, K., Hashimoto, K., Oura, K., and Tokuda, K. (2012). Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

Takaki, S., Sawada, K., Hashimoto, K., Oura, K., and Tokuda, K. (2013). Overview of NITECH HMM-based speech synthesis system for Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Tao, J., Li, Y., Pan, S., Zhang, M., Sun, H., and Wen, Z. (2009). The WISTON text-to-speech system for Blizzard Challenge 2009. In *Blizzard Challenge Workshop 2009*.

Tao, J., Pan, S., Li, Y., Wen, Z., and Wang, Y. (2010). The WISTON text to speech system for Blizzard Challenge 2010. In *Blizzard Challenge Workshop 2010*.

Tao, J., Yu, J., Huang, L., Liu, F., Jia, H., and Zhang, M. (2008). The wiston text to speech system for Blizzard 2008. In *Blizzard Challenge Workshop 2008*.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press, UK.

Toda, T., Kawai, H., Hirai, T., Ni, J., Nishizawa, N., Yamagishi, J., Tsuzaki, M., Tokuda, K., and Nakamura, S. (2006). Developing a test bed of English text-to-speech system

XIMERA for the Blizzard Challenge 2006. In *Blizzard Challenge Workshop 2006*.

Watts, O., Stan, A., Mamiya, Y., Suni, A., Burgos, J., and Montero, J. (2013). The Simple[4]All entry to the Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Weiss, C., Paulo, S., Figueira, L., and Oliveira, L. C. (2007). Blizzard entry: integrated voice building and synthesis for unit-selection TTS. In *Blizzard Challenge Workshop 2007*.

Wilhelms-Tricarico, R., Mottershead, B., Reichenbach, J., and Marple, G. (2012). The Lessac Technologies hybrid concatenated system for Blizzard Challenge 2012. In *Blizzard Challenge Workshop 2012*.

Wilhelms-Tricarico, R., Reichenbach, J., and Marple, G. (2013). The Lessac Technologies hybrid concatenated system for Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Wouters, J. (2007). SVOX participation in Blizzard 2007. In *Blizzard Challenge Workshop 2007*.

Yamagishi, J., Lincoln, M., King, S., Dines, J., Gibson, M., Tian, J., and Guan, Y. (2009). Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework. In *Blizzard Challenge Workshop 2009*.

Yamagishi, J. and Watts, O. (2010). The CSTR/EMIME HTS system for Blizzard Challenge. In *Blizzard Challenge Workshop 2010*.

Yamagishi, J., Zen, H., Toda, T., and Tokuda, K. (2007). Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the Blizzard Challenge 2007. In *Blizzard Challenge Workshop 2007*.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Blizzard Challenge Workshop 2008*.

Yang, J.-H., Zhao, Z.-W., Jiang, Y., Hu, G.-P., and Wu, X.-R. (2006). Multi-tier non-uniform unit selection for corpus-based speech synthesis. In *Blizzard Challenge Workshop 2006*.

Yu, Y., Zhu, F., Li, X., Liu, Y., Zou, J., Yang, Y., Yang, G., Fan, Z., and Wu, X. (2013). Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013. In *Blizzard Challenge Workshop 2013*.

Zen, H. and Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Blizzard Challenge Workshop 2005 (special session of Interspeech 2005)*.

Zen, H., Toda, T., and Tokuda, K. (2006). The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. In *Blizzard Challenge Workshop 2006*.

Zhang, B., Alhonen, J., Guan, Y., and Tian, J. (2010). Multilingual TTS system of Nokia entry for Blizzard 2010. In *Blizzard Challenge Workshop 2010*.

Zhang, Z., Xian, X., Luo, L., and Wu, X. (2009). PKU Mandarin speech synthesis system for Blizzard 2009. In *Blizzard Challenge Workshop 2009*.