Chapter 2
Challenges in Speech Synthesis

David Suendermann, Harald Höge, and Alan Black

2.1 Introduction

Similar to other speech- and language-processing disciplines such as speech recognition or machine translation, speech synthesis, the artificial production of human-like speech, has become very powerful over the last 10 years. This is not only due to the following reasons:

- Extensive scientific work rendered by hundreds of speech synthesis researchers worldwide.
- Ever-growing computational capacity. Approaches like unit selection require a significant processor load to be efficient and real-time able.
- Also, the more speech data is available and the higher its resolution is, the better the achieved quality can be. This, however, requires huge memory capacities able to provide random access to the speech data involved. E.g., for the speech data of the European project TC-Star, 10 h of high-fidelity speech with a sampling rate of 96 kHz/24 bit was used [10]. Even nowadays, only few computers are able to hold the whole speech data in their memory. It comprises almost 10 GB.
- Last but not least, as often in the history of science, money plays a key role for the development of certain research fields. In the speech synthesis domain, extensive financial resources are required in particular for generating the speech resources. The rigorous process of producing high-quality recordings of a synthesis voice requires several steps of speaker selection among professional subjects, special recording environments, careful utterance preparation, recording, transcription, annotation, labeling, and pitch tracking. Corpora like the aforementioned 10-h recording or even larger ones (see Section 2.3 for more examples) could only be provided in recent years since the trust in speech synthesis technology had significantly gained and funds could be raised.

D. Suendermann (✉)
SpeechCycle, Inc., 26 Broadway, 11th Floor, New York, NY, USA
e-mail: david@speechcycle.com

F. Chen, K. Jokinen (eds.), *Speech Technology*, DOI 10.1007/978-0-387-73819-2_2,

As quality and naturalness of speech synthesis significantly improved in the past 10 years and more and more techniques with supposedly superior quality and intelligibility were presented, a demand for a profound comparison between these techniques emerged. Inspired by its sister fields, the speech synthesis community examined the situation in speech recognition and machine translation research. Already early in the 1980s, a considerable number of speech recognition systems showed decent performance, but nobody could reliably tell which one was the best, since all of them had been trained and tested on different corpora and potentially used different performance metrics making it impossible to have a fair comparison. As a consequence, in 1982, Texas Instruments produced a carefully recorded US English multi-speaker corpus with more than 25,000 digit sequences which was to be used as test corpus for the competing parties [29]. At this time, one also agreed on a standard error measure (edit or Levenshtein distance [30]). Although the introduction of standard corpus and measure was a large step toward a fair comparison, several issues had not yet been resolved:

- Since every development team tested their recognizer against the standard corpus in their own laboratory, the numbers they published were not completely trustworthy.
- The same test corpus was used over and over again, and often the developers used it to actually tune their systems, which could result in a performance significantly higher than if the test corpus would have never seen before.
- The usage of a specific test corpus was entirely voluntary. Say, there were five well-recognized corpora, but a specific recognizer performed well only on one of them, there was no need to publish the worse results on the other corpora.

Fortunately, these problems could be resolved by introducing regular evaluation races which were performed by an independent institution. The very first of such competitions was the DARPA Resource Management project launched in 1987 which involved evaluation turns roughly every 6 months [41]. The National Institute of Standards in Telecommunications (NIST) served as independent evaluation institution. For every evaluation turn, a new test corpus was distributed making it impossible to tune on the test set. Also, once being registered for an evaluation, the respective party was expected to submit its recognition results—a withdrawal was considered a failure—and, furthermore, there was no way for the submitting party to predict the performance of its submission before the publication of all competitors' results.

In the mid-2000s, speech synthesis research tried to learn a lesson from the aforementioned developments in speech recognition and other areas by initiating a number of regular competitions meeting the above formulated criteria for objectivity of results including:

- independent evaluation institutions;
- standardized evaluation corpora changing from evaluation to evaluation;
- standardized evaluation metrics.

Being initiated only a few years ago, these competitions constitute a new trend in speech synthesis research. They build a platform joining the most important players in the field; and the frequently held evaluation workshops are discussion forums for the most recent speech synthesis technology. In this function, the competitions discussed in this chapter are the test bed for new trends in the field.

As speech synthesis competitions are based on historic experiences and, at the same time, markers for future trends in synthesis research, this chapter will first deal with the history of this discipline. In a second part, it will outline the most important state-of-the-art techniques and their representation in the scope of speech synthesis competitions.

2.2 Thousand Years of Speech Synthesis Research

2.2.1 From Middle Ages Over Enlightenment to Industrial Revolution: Mechanical Synthesizers

The history of speech synthesis, i.e., the artificial production of human-like speech, is presumably much longer than many of the readers might expect, and it is certainly the oldest speech processing discipline discussed in this book. Legends of talking "machines" go 1000 years back to Pope Sylvester II (950–1003 AD) who was supposed to possess a *brazen head* [12]. This kind of prophetic device was reputed to be able to answer any question [11]. In this spirit, it can be regarded to have been the very first dialog system including components of speech recognition, understanding, generation, and, last but not least, synthesis.

Indeed, referring to the brazen head of Sylvester II as the first automatic speech-processing device is as reasonable as calling Icarus' wings the first airplane. But as much as human's dream to fly eventually came true, not only magicians, but some centuries later well-reputed scientists like Isaac Newton addressed the production of artificial speech [28]. In 1665, the Fellow of the Royal Society carried out experiments pouring beer into bottles of different shapes and sizes and blowing into them producing vowel-like sounds. This observation was exploited 300 years later for the development of linear predictive coding [33], one of the most important speech analysis and synthesis techniques.

Before the era of linear predictive coding, there were several attempts to build mechanical devices able to produce human-like speech, as for instance the one of Wolfgang von Kempelen [47]. He was a scientist in the service of Empress Maria Theresa in Vienna at the end of the 18th century. Von Kempelen is considered one of the first experimental phoneticians who built several outstanding devices as the *Turk* [44], a supposedly chess-playing machine which later turned out to be a hoax, and his speaking machine [14]. The latter was an apparatus composed of a bellow representing the lungs, a rubber tube for the mouth, and a wooden extension being the nose (for a construction drawing, see Fig. 2.1). By means of two levers controlling the resonance characteristics, a complete set of a language's sounds could

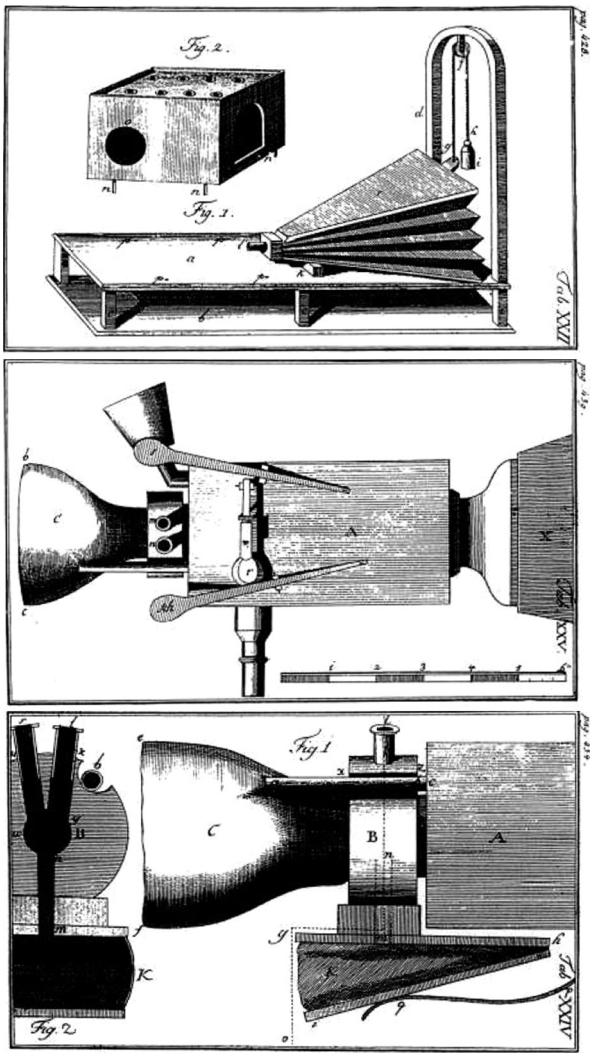

Fig. 2.1 The main components of von Kempelen's speaking machine

be produced. Von Kempelen stated that he successfully produced whole words and sentences in languages such as Latin and French.

In the following years, other inventors produced similar and advanced devices for human-like speech generation, e.g.

- Erasmus Darwin's (the grandfather of the evolutionist) speech synthesizer (1806) [13];
- Joseph Faber's *Euphonia* (1844) [31]; and
- Alexander Graham Bell's (the father of telephony) physical working model of the human vocal tract (end of the 19th century) [3, 17].

2.2.2 The 20th Century: Electronic Synthesizers

The electronic age opened new horizons to speech synthesis technology. After the development of electrical oscillators, filters, amplifiers, loudspeakers, etc., it was possible to generate and control sounds much easier than with mechanical devices whose sound production always was limited to its physical dimensions.

As early as in the year 1936, the UK Telephone Company manufactured an automatic clock, the first instance of practical text-to-speech synthesis [21]. Partial utterances such as numbers and time expressions were stored on a glass disk and were concatenated to form complete sentences.

Only 3 years later, another significant attempt to create a speaking machine (human operated as von Kempelen's) based on electronic sound generation was developed at the Bell Labs (named after the aforementioned Alexander Graham Bell). Homer Dudley's *VODER* (Voice Operated recorDER) was first presented at the 1939 World Fair in New York. It was a keyboard-controlled apparatus composed of a "vocal tract" with 10 parallel band-pass filters spanning the speech frequency range, and an excitation which could be either unvoiced or voiced. Unvoiced sounds were produced by means of a random noise generator; voiced sounds came from a relaxation generator whose fundamental frequency was controlled through a foot pedal. Keyboard and pedal clearly remind of the design of a musical keyboard, and, interestingly, Homer Dudley's visit at Bonn University in 1948 (the year before Bonn became capital of West Germany) inspired Professor Werner Meyer-Eppler to apply synthesis devices such as the VODER to music composition pioneering the Elektronische Musik movement [27]. In this sense, speech and music synthesizers have the same roots.

All of the attempts to this date had been based on analog signal processing. The development of pulse code modulation (PCM) to transform analog and digital speech representations into each other in the beginning of the 1940s built the foundation of digital speech processing. The very first application of transmitting digitized speech was used during World War II in the vocoder encryption equipment of SIGSALY, the secure speech system connecting London and the Pentagon [6].

But PCM was not the only ground-breaking novelty in the mid-20th century. Until the end of the 1950s, experimental research on electronic speech synthesis was focused on specialized devices which were hard-coded and bound to their original purpose. The development of computers, however, gave a lot more flexibility, and as soon as they became strong enough to store and process digitized speech, they were extensively applied to speech-processing tasks.

Again at Bell Labs, probably the first time in history, John L. Kelly used a computer (IBM 704) to create synthetic speech—and not only speech but even a lilting voice singing the song *Bicycle Built for Two*. This incidence later inspired John's friend (and namesake) John Pierce to use this (or a similar) sound sample at the climactic scene of the screenplay for *2001: A Space Odyssey* [48].

At this time, the sound of digital speech synthesis was far from natural, but the clearly recognizable computer voice had its own charm, and synthesis capability was integrated into a number of rather successful electronic devices and computer systems. Very popular was the electronic toy *Speak & Spell*, since 1978

manufactured by Texas Instruments [15]. It contained a single-chip speech synthesizer, the TI TMC0280, based on 10th order linear predictive coding. *Speak & Spell* had a small membrane keyboard and a vacuum fluorescent display and was designed to help young children to become literate, learn the alphabet, and to spell. Again, this synthesizing device played a prominent role in the movies: In Steven Spielberg's motion picture *E.T. the Extra-Terrestrial*, it served as a key component of the alien's home-built interstellar communicator.

Speech synthesis as software or as an integral part of the operating system was introduced in the beginning of the 1980s on computers such as Apple Macintosh or Commodore Amiga.[1] The ever-increasing power of computers witnessed over the last 20 years allowed for the emerging of a wide variety of synthesis techniques whose development and refinement is continued until these days. In the scope of speech synthesis competitions—focus of the remainder of this chapter—many of these techniques are extensively discussed; the most important ones will be revisited in the following section.

2.3 The Many Hats of Speech Synthesis Challenges

2.3.1 Evaluation, Standardization, and Scientific Exchange

In the first years of the new millennium, several groups in Europe, the United States, and Asia met to discuss opportunities for competitions in speech synthesis. Focus elements of all these efforts were:

- the release of *standard speech databases* that would be used by every participant of these challenges to isolate the impact of the speech database from the core synthesis technology,
- the usage of *standard evaluation metrics* to have a fair comparison between the partners and provide an easy measure to express the performance of a system in absolute terms, and
- the involvement of an *independent evaluation party* that would prepare, send out, collect, and evaluate speech synthesis utterances, build an evaluation framework, hire and supervise subjects, and finally report on the evaluation results.

The very first consortium of internationally renowned partners setting up such a framework was the project TC-Star (Technology and Corpora for Speech-to-Speech Translation) funded by the European Commission. Project members were 12 prestigious research institutions from industry and academia. The project's main goal was to significantly reduce the gap between human and machine translation performance. The 36-month project starting on April 1, 2004 was to support basic research in the areas of speech recognition, machine translation, and speech synthesis in the domain of parliamentarian and other political speeches.

[1] The first author of this chapter used an Amiga 500 computer built in the mid-1980s to synthesize voices for his 1997 drum and bass release *Digital Emperator: Out of O2* [51].

Although the project proposal dates back to as early as 2002 [20], the very first speech synthesis evaluation was conducted only in September 2005. This long delay was not only due to the project being launched in spring 2004, but also due to the very ambitious goal of releasing five multilingual voices for the world's three most frequently spoken languages (Mandarin, English, Spanish). Each of these five voices included 10 h of speech recorded in highly clean studio environments with 96 kHz/24 bit recording precision making the process slow and expensive [10].

The first call for participation to another large-scale initiative on speech synthesis competition also dates back to the year 2004. The Blizzard challenge was specifically founded to compare technology for corpus-based speech synthesis [8]. In doing so, a database that was intentionally designed without the use of copyrighted material was prepared at the Carnegie Mellon University and released free of charge to all participants of the challenge. The first release of the CMU Arctic databases (June 2004) consisted of one female and one male US English voice uttering about 1200 phonetically balanced utterances with a total duration of around 1.5 h each [26]. These corpora were recorded under 16 kHz/16 bit studio conditions and, compared to TC-Star, much faster to record and compile. Consequently, the time span from starting the database preparation and sending to the participants, to the actual evaluation was altogether about a year and a half—the first Blizzard Challenge took place in January 2005, 8 months before the first TC-Star evaluation.

Not only did these competitions differ in the nature of the databases they used, but also the third aforementioned focus of these challenges, the evaluation framework, was treated differently in both instances.

On the one hand, TC-Star took the sister field speech recognition as driving example when using a well-recognized international standardization institution to carry out an independent evaluation as NIST did in the DARPA speech recognition competitions, cf. Section 2.1. The Evaluation and Language Resources Distribution Agency (ELDA, located in Paris, France) took over the responsibility for performing the evaluations including hiring subjects for the subjective assessment and carrying out objective measures if applicable.

On the other hand, the Blizzard challenge relied on in-house resources to perform the evaluation. This was mainly driven by lack of funding for a large independent and human-driven evaluation and by the conviction that, the more subjects participate in the rating, the more trustful the outcomes will be. Also, the evaluation approach itself and the selection of most appropriate subjects were subject to the curiosity of the involved researchers, since it had not yet been fundamentally investigated, to which extend the nature of the subjects is reflected in the outcomes of evaluations. At any rate, in the case of the Blizzard challenge 2005, the set of subjects was composed of:

- *Speech experts.* Every participant in the challenge was to provide 10 experts of its group to contribute.
- *Volunteers.* Random people using the web interface.
- *Undergraduate students.* These people were paid for task completion.

Also early in 2004, undoubtedly the initiation year of speech synthesis challenges, a third consortium was founded to bring together international experts of the speech synthesis community (mostly from Europe and Asia). Main goal of the European Center of Excellence in Speech Synthesis (ECESS) is the standardization, modularization, and evaluation of speech synthesis technology.

This consortium defined a standard text-to-speech system to consist of three main modules: the text processing, the prosody generation, and the acoustic synthesis. Each participant agrees on delivering an executable version of at least one of these three modules following a well-defined interface terminology [42]. Text-to-speech synthesis is considered a plug-and-play system where every partner can focus his research on one module making use of highly performing modules of other parties. To this end, module-wise evaluations are carried out in regular intervals to rate the performance of the evaluating modules.

All three, TC-Star, Blizzard Challenge, and ECESS, have a lot in common. They must not be considered as competing platforms where progress is gained only by trying to best the other participants. Instead, they are scientific forums that are driven by shared resources (such as the corpora in the Blizzard Challenges, or technological modules in ECESS), the grand effort of standardization (corpora, evaluation metrics, module interfaces, etc.), and an intensive dialog in the scope of regular workshops with scientific presentations and proceedings compiling peer-reviewed publications of the respective consortia. Such workshops were held every year during the life cycle of TC-Star. Blizzard Challenges are focus of annual workshops mostly organized as satellite events of major speech conferences such as Interspeech. ECESS workshops are held at least semi-annually since February 2004 in diverse places mostly in Europe. These events quickly became important venues for the major players of the field—the sheer numbers of participants suggests which important role these challenges and their by-products are playing in the speech synthesis community—see Table 2.1.

Table 2.1 Evolution of number of participants in speech synthesis challenges

	TC-Star [10, 36, 37]	Blizzard Challenge [8, 5, 18]	ECESS [42, 23]
2004	3	-	6
2005	3	6	13
2006	10	14	16
2007	6	16	17

2.3.2 Techniques of Speech Synthesis

This section is to give a brief overview about the most popular and successful techniques covered in evaluation campaigns and scientific workshops of speech synthesis challenges.

2.3.2.1 Concatenative Synthesis

As aforementioned, the emergence of powerful computers was main driver for the development of electronic speech synthesis. Mass storage devices, large memory, and high CPU speed were conditions of a synthesis paradigm based on concatenation of real speech segments. The speech of a voice talent is recorded and stored after digitizing. At synthesis time, the computer produces the target speech by cutting out segments from the recording and concatenating them according to a certain scheme.

This principle allows for very simple applications such as a talking clock where an arbitrary time expression is composed of a number of prerecorded utterances. You can imagine that one records the utterances "the time is," "a.m.," "p.m.," and the numbers from 1 to 59 and concatenates them according to the current time. This is indeed a straightforward and practical solution that was already used in the semi-mechanical talking clock by the UK Telephone Company mentioned earlier in this chapter.

Going from a very specialized application to general-purpose text-to-speech synthesis, the above described solution is not feasible anymore. We do not know in advance which words and phrases will be synthesized. Even when one would record every single word in the vocabulary of a language (which is theoretically possible), word-by-word concatenation would sound very fragmented and artificial. In natural speech, words are often pronounced in clusters (without pauses between them), and, most importantly, natural speech features a sentence prosody, i.e., depending on position, role in the sentence, sentence type, etc., word stress, pitch, and duration are altered.

In order to produce synthetic speech following these principles, Hunt and Black [22] proposed the *unit selection* technique. Basically, unit selection allows for using a speech database of arbitrary textual content, phonetically labeled. When a sentence is to be synthesized, the phonetic transcription is produced, and the sentence prosody is predicted. Now, units (of arbitrary length, i.e., number of consecutive phonemes) are selected from the database such that the phonetic transcription matches the target. In doing so, also the prosody of the selected units is considered to be closest as possible to the predicted target prosody. This optimization is carried out as a large search through the space of exponentially many possible selections of units matching the target transcription and is only tractable by applying dynamic programming [4] and beam search [19]. To account for prosody artifacts (pitch jumps, etc.), often, signal processing is applied to smooth concatenation points.

Over the years after the introduction of unit selection, there have been major improvements to optimize the quality of the output speech such as:

- paying extreme attention to a superior quality of the speech recordings to make them consistent, having accurate phonetic labels and pitch marks, highest possible SNR, etc. (see the careful recording specifications for TC-Star [10]),
- reducing involved signal processing as much as possible by producing a speech corpus that as best as possible reflects the textual and prosodic characteristics of the target speech [7],

- enlarging the corpus, since this increases the probability that nice units will be found in the database (see, e.g., the 16-h corpus ATRECCS provided for the Blizzard Challenge 2007 [39]),
- improving prosody models to produce more natural target speech [43].

Unit selection is definitely the most popular text-to-speech synthesis technique to date as for instance the submissions to the Blizzard Challenge 2007 show: 14 out of 15 publications at the evaluation workshop dealt with unit selection synthesis.

2.3.2.2 HMM-Based Synthesis

Hidden Markov models (HMMs) [1] were used since the mid-1970s as a successful approach to speech recognition [45]. The idea is that constant time frames of the PCM-encoded acoustic signal are converted into feature vectors representing the most important spectral characteristics of the signal. Sequences of feature vectors are correlated to phoneme sequences mapped to the underlying text through a pronunciation lexicon. The HMM models the statistical behavior of the signal and is trained based on a significant amount of transcribed speech data. In recognition phase, it allows for estimating the probability of a feature vector sequence given a word sequence. This probability is referred to as *acoustic model*.

Besides, HMMs allow for estimating the probability of a word sequence, called the *language model*. In this case, the training is carried out using a huge amount of training text.

The combination of acoustic and language model according to Bayes' theorem [2] produces the probability of a word sequence given a feature vector sequence. By trying all possible word sequences and maximizing this probability, the recognition hypothesis is produced.

In speech synthesis, this process has to be reversed [35]. Here, we are given the word sequence and search for the optimal sequence of feature vectors. Since the word sequence is known, we do not need the language model in this case. For applying the acoustic model, we first have to convert the word sequence into a phonetic sequence using pronunciation lexicon or (in case of an unknown word) grapheme-to-phoneme conversion (as also done in other synthesis techniques such as unit selection). The phoneme sequence is then applied to the acoustic model emitting the most probable feature sequence as described, e.g., in [9].

In contrast to the application to speech recognition, in synthesis, the consideration of spectral features alone is not sufficient, since one has to produce a wave form to get audible speech. This is done by generating a voiced or unvoiced excitation and filtering it by means of the spectral features on a frame-by-frame basis. This procedure is very similar to the VODER discussed in Section 2.2. To overcome the limitation to just two different excitation types resulting in rather synthetic and muffled speech, there are several approaches to more sophisticated excitation modeling like, for instance, the harmonic + noise model [50], residual prediction [52], or mixed excitation with state-dependent filters [32].

Convincing advantage of HMM-based speech synthesis is its small footprint. As this technique only requires the model parameters to be stored (and not the speech data itself), its size is usually limited to a few megabytes. In contrast, the above-mentioned 16-h synthesis corpus ATRECCS, sampled at 48 kHz and a final 16 bit precision requires more than 5 GB of storage.

In addition, HMM-based synthesis requires less computation than unit selection. This is because the latter searches a huge space becoming larger and larger as more speech data are available. Real-time ability can be an issue with such systems, whereas HMM-based synthesis operates at much lower expense.

Last but not least, HMM-based synthesis has the potential to produce high-quality speech. This is mainly because HMMs produce continuous spectral contours as opposed to unit selection synthesis where we may face inconsistencies at the concatenation points resulting in audible speech artifacts. Blizzard Challenge 2005 delivered the proof that HMM-based synthesis is even able to outperform the well-established unit selection technique, as the contribution [55] achieved the best results in both speech quality and intelligibility. This, however, was partially because the training corpus (the Arctic database as introduced above) contained the relative small amount of 1.5 h of speech. Unit selection synthesis may face unit sparseness problems with too small databases lacking appropriate data for the target utterances: In the Blizzard Challenge 2006 which provided a 5-h corpus, the best system was based on unit selection synthesis [25].

2.3.2.3 Voice Conversion

In the above sections, we emphasized the importance of large and carefully designed corpora for the production of high-quality speech synthesizers. The compilation of corpora such as those built for TC-Star or the ATRECCS database incorporates a significant time effort as well as high monetary expenses. This explains that, usually, synthesizers come along with at most two or three voices for a given language constituting a lack of variety and flexibility for the customer of speech synthesis technology.

Voice conversion is a solution to this problem. It is a technology that allows for rapidly transforming a source voice into a target voice [38]. Statistical approaches to voice conversion discussed since the mid-1990s [49, 24] are the most popular ones. These approaches link parallel source and target speech by means of spectral feature vectors and generate a joint statistical model, usually a Gaussian mixture model (GMM), representing the relation between both voices. In this context, *parallel speech* means that both source and target speaker uttered the very same text preferably with a similar timing pattern. A one-to-one speech frame mapping is then produced by alignment techniques such as dynamic time warping [46] or HMM-based forced alignment [54]. This approach is called *text dependent*, since it requires the target speaker to utter the same text as the source speaker. In contrast, *text-independent* techniques provide flexibility regarding the target speech as required if voice conversion is to be applied on earlier recorded databases, including the case

that both voices use different *languages*. In this case, one speaks of *cross-language* voice conversion [34].

The application of voice conversion to speech synthesis was one of the focuses of the TC-Star project dedicating entire workshop sessions to this topic, e.g., the 2006 workshop in Barcelona dealt with GMM-based [40], text-independent [16], as well as cross-language voice conversion [53].

2.4 Conclusion

Competitions (or challenges) in speech synthesis emerging only few years ago are playfields for research and industry playing several roles at once.

- Originally, they were mainly intended to serve as a platform where different synthesis techniques could be compared.
- To design this comparison as objective as possible, standard corpora, standard evaluation measures, and a standard evaluation framework were proposed. In this way, and particularly as the number of participants grew significantly, challenges became forums for standardization of corpora, evaluation criteria, and module interfaces.
- Challenge-related workshops, conferences, and meetings evolved from pure result-reporting events to scientific venues with talks, poster presentations, and published proceedings, hence a platform for the exchange of ideas.
- They serve challenges not only for the exchange of ideas but also for the exchange of technology. Special agreements between participants led to a publication of significant parts of used resources and software. Corpora are made available for download in the Internet; source codes and binaries of entire speech synthesizers, modules, tools, and evaluation kits are distributed.

The authors believe that a major part of future speech synthesis research will be carried out in the scope of such challenges leading to a boost of quality and applicability of speech synthesis solutions and a strengthening of the whole research field.

References

1. Baum, L., Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Statistics, 37, 1554–1563.
2. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Philos. Trans. Roy. Soc. Lond., 53, 370–418.
3. Bell, A. (1922). Prehistoric telephone days. Natl. Geographic Mag., 41, 223–242.
4. Bellmann, R. (1957). Dynamic Programming. Princeton University Press, Princeton, USA.
5. Bennett, C., Black, A. (2006). The Blizzard Challenge 2006. In: Blizzard Challenge Workshop, Pittsburgh, USA.

6. Bennett, W. (1983). Secret telephony as a historical example of spread-spectrum communications. IEEE Trans. Commun., 31(1), 98–104.
7. Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A. (2006). The AT&T Next-Gen TTS system. In: Proc. TC-Star Workshop, Barcelona, Spain.
8. Black, A., Tokuda, K. (2005). Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In: Proc. Interspeech, Lisbon, Portugal.
9. Black, A., Zen, H., Tokuda, K. (2007). Statistical parametric synthesis. In: Proc. ICASSP, Honululu, USA.
10. Bonafonte, A., Höge, H., Tropf, H., Moreno, A., v. d. Heuvel, H., Sündermann, D., Ziegenhain, U., Pérez, J., Kiss, I. (2005). TC-Star: Specifications of language resources for speech synthesis. Technical Report.
11. Butler, E. (1948). The Myth of the Magus. Cambridge University Press, Cambridge, UK.
12. Darlington, O. (1947). Gerbert, the teacher. Am. Historical Rev., 52, 456–476.
13. Darwin, E. (1806). The Temple of Nature. J. Johnson, London, UK.
14. Dudley, H., Tarnoczy, T. (1950). The speaking machine of Wolfgang von Kempelen. J. Acoust. Soc. Am., 22(2), 151–166.
15. Dutoit, T. (1997). An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht, Netherlands.
16. Duxans, H., Erro, D., Pérez, J., Diego, F., Bonafonte, A., Moreno, A. (2006). Voice conversion of non-aligned data using unit selection. In: Proc. TC-Star Workshop, Barcelona, Spain.
17. Flanagan, J. (1972). Voices of men and machines. J. Acoust. Soc. Am., 51, 1375–1387.
18. Fraser, M. King, S. (2007). The Blizzard challenge 2007. In: Proc. ISCA Workshop on Speech Synthesis, Bonn, Germany.
19. Hand, D., Smyth, P., Mannila, H. (2001). Principles of Data Mining. MIT Press, Cambridge, USA.
20. Höge, H. (2002). Project proposal TC-STAR – Make speech to speech translation real. In: Proc. LREC, Las Palmas, Spain.
21. Holmes, J., Holmes, W. (2001). Speech Synthesis and Recognition. Taylor and Francis, London, UK.
22. Hunt, A., Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP, Atlanta, USA.
23. Kacic, Z. (2004–2007). Proc. 11th–14th Int. Workshops on Advances in Speech Technology. University of Maribor, Maribor, Slovenia.
24. Kain, A., Macon, M. (1998). Spectral voice conversion for text-to-speech synthesis. In: Proc. ICASSP, Seattle, USA.
25. Kaszczuk, M., Osowski, L. (2006). Evaluating Ivona speech synthesis system for Blizzard Challenge 2006. In: Blizzard Challenge Workshop, Pittsburgh, USA.
26. Kominek, J., Black, A. (2004). The CMU arctic speech databases. In: Proc. ISCA Workshop on Speech Synthesis, Pittsburgh, USA.
27. Kostelanetz, R. (1996). Classic Essays on Twentieth-Century Music. Schirmer Books, New York, USA.
28. Ladefoged, P. (1998). A Course in Phonetics. Harcourt Brace Jovanovich, New York, USA.
29. Leonard, R., Doddington, G. (1982). A Speaker-Independent Connected-Digit Database. Texas Instruments, Dallas, USA.
30. Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys. Dokl., 10, 707–710.
31. Lindsay, D. (1997). Talking head. Am. Heritage Invention Technol., 13(1), 57–63.
32. Maia, R., Toda, T., Zen, H., Nankaku, Y., Tokuda, K. (2007). An excitation model for HMM-based speech synthesis based on residual modeling. In: Proc. ISCA Workshop on Speech Synthesis, Bonn, Germany.
33. Markel, J., Gray, A. (1976). Linear Prediction of Speech. Springer, New York, USA.
34. Mashimo, M., Toda, T., Shikano, K., Campbell, N. (2001). Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In: Proc. Eurospeech, Aalborg, Denmark.

35. Masuko, T. (2002). HMM-based speech synthesis and its applications. PhD thesis, Tokyo Institute of Technology, Tokyo, Japan.
36. Mostefa, D., Garcia, M.-N., Hamon, O., Moreau, N. (2006). TC-Star: D16 Evaluation Report. Technical Report.
37. Mostefa, D., Hamon, O., Moreau, N., Choukri, K. (2007). TC-Star: D30 Evaluation Report. Technical Report.
38. Moulines, E. and Sagisaka, Y. (1995). Voice conversion: State of the art and perspectives. Speech Commun., 16(2), 125–126.
39. Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R., Nakamura, S. (2007). ATRECSS – ATR English speech corpus for speech synthesis. In: Proc. ISCA Workshop on Speech Synthesis, Bonn, Germany.
40. Nurminen, J., Popa, V., Tian, J., Tang, Y., Kiss, I. (2006). A parametric approach for voice conversion. In: Proc. TC-Star Workshop, Barcelona, Spain.
41. Pallet, D. (1987). Test procedures for the March 1987 DARPA Benchmark Tests. In: Proc. DARPA Speech Recognition Workshop, San Diego, USA.
42. Pérez, J., Bonafonte, A., Hain, H.-U., Keller, E., Breuer, S., Tian, J. (2006). ECESS inter-module interface specification for speech synthesis. In: Proc. LREC, Genoa, Italy.
43. Pfitzinger, H. (2006). Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In: Proc. Speech Prosody, Dresden, Germany.
44. Poe, E. (1836). Maelzel's Chess Player. Southern Literary Messenger, 2(5), 318–326.
45. Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77(2), 257–286.
46. Rabiner, L., Rosenberg, A., Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. IEEE Trans. Acoustics, Speech Signal Process., 26(6), 575–582.
47. Ritter von Kempelen, W. (1791). Mechanismus der menschlichen Sprache nebst der Beschreibung einer sprechenden Maschine. J. V. Degen, Vienna, Austria.
48. Stork, D. (1996). HAL's Legacy: 2001's Computer as Dream and Reality. MIT Press, Cambridge, USA.
49. Stylianou, Y., Cappé, O., Moulines, E. (1995). Statistical methods for voice quality transformation. In: Proc. Eurospeech, Madrid, Spain.
50. Stylianou, Y., Laroche, J., Moulines, E. (1995). High-quality speech modification based on a harmonic + noise model. In: Proc. Eurospeech, Madrid, Spain.
51. Suendermann, D., Raeder, H. (1997). Digital Emperator: Out of O2. d.l.h.-productions, Cologne, Germany.
52. Sündermann, D., Bonafonte, A., Ney, H., Höge, H. (2005). A study on residual prediction techniques for voice conversion. In: Proc. ICASSP, Philadelphia, USA.
53. Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J. (2006). TC-Star: Cross-language voice conversion revisited. In: Proc. TC-Star Workshop, Barcelona, Spain.
54. Young, S., Woodland, P., Byrne, W. (1993). The HTK Book, Version 1.5. Cambridge University Press, Cambridge, UK.
55. Zen, H., Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In: Proc. Interspeech, Lisbon, Portugal.

http://www.springer.com/978-0-387-73818-5

Speech Technology
Theory and Applications
Chen, F.; Jokinen, K. (Eds.)
2010, XXVII, 331 p. 188 illus., 23 illus. in color.,
Hardcover
ISBN: 978-0-387-73818-5