Speech Synthesis: Text-To-Speech Conversion and Artificial Voices

Jürgen Trouvain and Bernd Möbius

Contents

Abstract

The artificial generation of speech has fascinated mankind since ancient times. The robotic-sounding artificial voices from the last century are nowadays replaced with more naturally sounding voices based on pre-recorded human speech. Significant progress in data processing led to qualitative leaps in intelligibility and naturalness. Apart from sizable data of the voice donor, a fully fledged text-to-speech (TTS) synthesizer requires further linguistic resources and components of natural language processing including dictionaries with information on pronunciation and word prosody, morphological structure, and parts-of-speech but also procedures for automatic chunking texts in smaller parts, or morpho-syntactic parsing. TTS technology can be used in many different application domains, for instance, as a communicative aid for those who cannot speak and those who cannot see and in situations characterized as "hands busy, eyes busy" often as a part of spoken dialog systems. One remaining big challenge

J. Trouvain (✉) · B. Möbius
Department of Language Science and Technology, Saarland University, Saarbrücken, Germany
e-mail: trouvain@coli.uni-saarland.de; moebius@coli.uni-saarland.de

S. D. Brunn, R. Kehrein (eds.), *Handbook of the Changing World Language Map*,
https://doi.org/10.1007/978-3-319-73400-2_168-1

is evaluation of the quality of synthetic speech output and its appropriateness for the needs of the user. There are also promising developments in speech synthesis that go beyond the pure acoustic channel. Multimodal synthesis includes the visual channel, e.g., in talking heads, whereas silent-speech interfaces and brain-to-speech conversion convert articulatory gestures and brain waves, respectively, to spoken output. Although there has been much progress in quality in the last decade, often achieved by processing enormous amounts of data, TTS today is available only for relatively few languages (probably fewer than 50 with a dominance of English). Thus, a major task will be to find or create linguistic resources and make them available for more languages and language varieties.

Keywords
Text-to-speech · Speech synthesis · Artificial voices · Communicative aids · Natural language processing

Introduction

Synthesis of speech with the help of machines can mean various things on different levels. It can mean that the speech acoustics of the voice is artificially generated through hardware or software. In those cases, the human voice is completely imitated mechanically or electronically. In other approaches the synthesized speech is based on recordings of a human speaker. These recordings of natural speech are subsequently processed in various ways with the aim to generate new speech signals. In this sense the voice of the generated speech is an artificial voice, too.

Speech synthesis can also mean the conversion from text to speech (or TTS in short). For this endeavor not only an artificial voice is needed but also some knowledge about the pronunciation and the prosody of the text to be generated. Thus, speech synthesis can be seen as consisting of a linguistic processing part and a voice generation part.

There are many motivations and reasons why people generate synthetic speech and listen to synthetic voices. Blind persons can use TTS technology as a reading device. Individuals who are unable to produce speech with their own bodies can use TTS as a speech prosthesis. A navigation device in a car uses synthetic speech, likewise dialog systems and personal assistants. More details can be found in the section on applications of TTS technology.

The range of applications makes it clear that there are many situations where technical support via TTS technology is useful for communication with machines and in some cases substantially improves the life quality of people. However, usage of TTS devices today is still rather limited. For instance, the technology is available only for a minority of users worldwide because there are no or few linguistic and technical resources that are needed for TTS generation.

The artificial generation of speech has fascinated mankind since ancient times. The usual setting was that of an artificial talking head that produced speech in some “magical” way, either by a hidden human speaker (i.e., no automatic speech

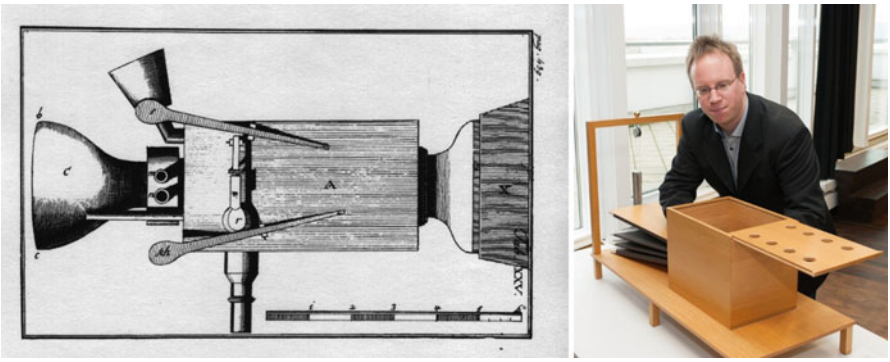

Fig. 1 Kempelen's drawing of the "inner life" of his speaking machine (left) and researcher Fabian Brackhane who demonstrates how to "play" a replica of the machine (right)

generation) or generated by a mechanical system with preformed sentences (automation but playback only). The beginning of a functioning mechanical speech synthesis can be dated back to the end of the eighteenth century. With his speaking machine, von Kempelen (1734–1804) was able to mechanically generate short utterances in different languages by using bellows, a "voice" consisting of a small vibrating ivory plate, and several resonance bodies made of wood, metal, and rubber (Fig. 1). The artificial voice "out of the box" was recognized by many listeners as that of a toddler's voice (von Kempelen 2017).

With the beginning of the twentieth century, the first steps of electronical speech synthesis were taken. The Voder (Voice Operation DEmonstratoR) was presented to the general public at the New York World's Fair in 1939, and its artificial speech was transmitted by radio to listeners across the USA. It consisted of electronic oscillators and a noise generator for the sound sources and resonator filters for the vocal tract. A skilled operator manually controlled pressure-sensitive keys for the production of vowels and consonants and a foot pedal for the fundamental frequency (pitch) (Fig. 2).

Modelling the acoustics of speech remained the predominant method of synthesis until the 1980s. The resulting robotic-sounding artificial voices were subsequently replaced with more naturally sounding voices based on the concatenation of segments of read speech recorded by human speakers.

Exponential increases in data storage and processing capacities in the 1990s and later made it possible to use far larger datasets of natural speech. This has led to qualitative leaps in intelligibility and naturalness at least for languages with sufficiently large speech datasets.

Text-to-Speech System Architecture

A typical architecture of a text-to-speech synthesis system consists of two components, one being concerned with symbol processing and the other with speech signal generation (Fig. 3) (for overviews see Dutoit 1997; Sproat 1998; Taylor 2009).

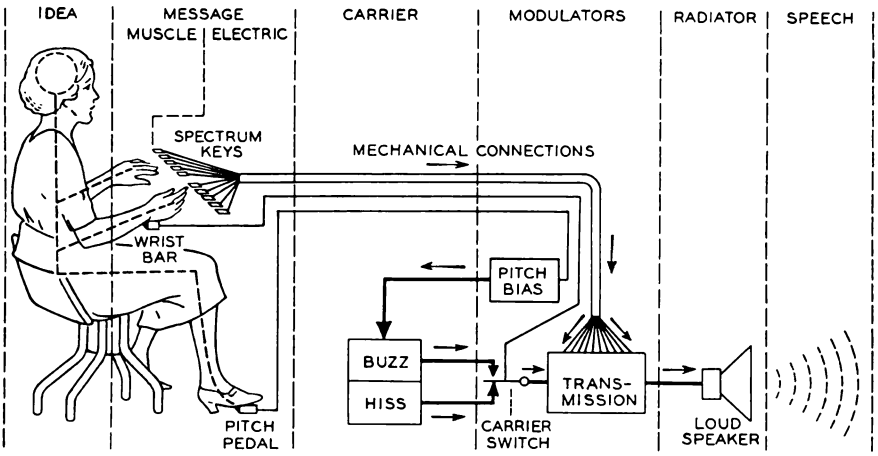

Fig. 2 Schematic view of the circuit of the Voder, the first electronical speech synthesizer (Dudley 1940)

Fig. 3 The general architecture of a text-to-speech synthesis system, consisting of two components, one being concerned with text analysis (in green), the other with speech signal generation (in blue)

These two components each consist of several modules handling specific tasks in text-to-speech conversion. The architecture is usually designed as a pipeline of modules. Each module takes as its input the output of the preceding module, performs its own specific task, and hands the modified representation over to the subsequent module. The pipeline is unidirectional, which entails that a mistake in an early step is difficult to remedy at a later stage of processing and typically triggers further mistakes. Even though alternative, more interactive architectures have been proposed, the pipeline architecture is still the prevalent one.

In the following sections, we will describe the text analysis (Fig. 4) and speech synthesis components (Fig. 5) in more detail.

Text Analysis

Modules of the text analysis component are concerned with inferring the linguistic structure of the input text (see Fig. 4). The first module takes the text to be synthesized and performs a *text normalization* (or regularization) task. Consider the text in (1a):

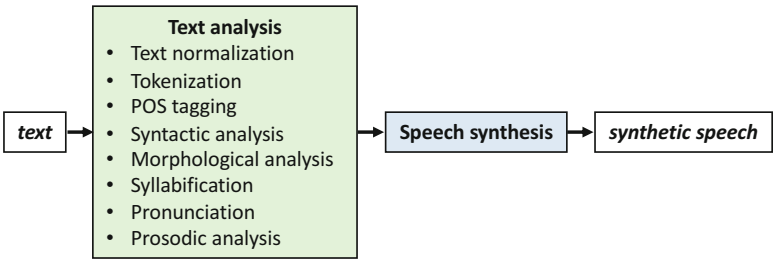

Fig. 4 Modules of the linguistic text analysis component (in green)

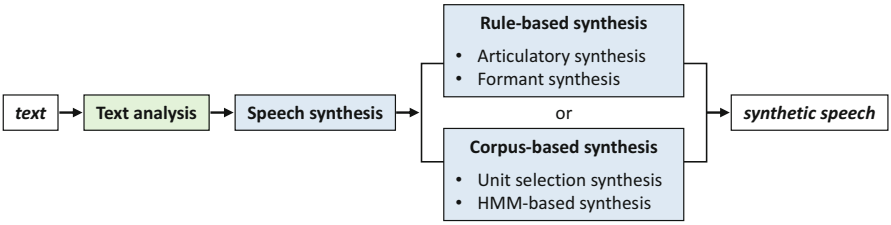

Fig. 5 Different strategies of generating a synthetic speech signal (in blue)

(1a) *Last weekend Dr. Smith read that the EU's recommendation is to impose a speed limit of 55 mph.*

(1b) [begin-of-sentence] last weekend [minor-break] doctor smith read [pronounce-as-"red"] [minor-break] that the EU's[pronounce-as-"ee YOOZE"] recommendation is to impose a speed limit of 55 miles per hour [major-break, end-of-sentence].

Text normalization converts items such as acronyms, abbreviations, and numeric expressions into regular word forms. In the text in (1a), this module will ideally disambiguate punctuation marks such as the period, whose first instance marks an abbreviation and whose second instance marks the end of the sentence; mark the acronym "EU" to be spelled out letter by letter; expand the abbreviation "Dr."; and convert the alphanumeric expression "55 mph" into a sequence of words. End-of-sentence detection and word segmentation are tasks referred to as *tokenization*. For instance, end-of-sentence detection is trivial in Chinese because the Chinese writing system uses a unique symbol for this purpose, whereas the period has several other functions in English. On the other hand, words are separated by white space in English, whereas Chinese does not explicitly mark word boundaries in the text.

Subsequent modules tag the words in the text for their *parts-of-speech* (POS) and analyze the *syntactic structure*. The latter helps to insert minor breaks at locations suggested by the phrase structure. The former provides information about which words may be accented: usually nouns and verbs, and sometimes adjectives, are phonetically more prominent than grammatical function words. In the example in

(1a), the verb form "read" is ambiguous with respect to tense. Context analysis helps determine that past tense is the correct form here: present tense is ruled out because the verb must agree with the subject, a single person (Dr. Smith), which would require the third person singular present tense form "reads." This fine-grained grammatical analysis also disambiguates the pronunciation ([rɛd] rather than [rid]).

In many languages, the internal structure of *morphologically complex words* must be analyzed to infer the correct *pronunciation*. For instance, the German compound "Wachstube" is ambiguous and has different pronunciations and meanings, depending on its internal structure: "Wach+Stube" [vaxʃtuːbə] (*guardroom*) vs. "Wachs+Tube" [vakstuːbə] (*tube of wax*). The actual pronunciation of words is looked up in a built-in pronunciation dictionary or, failing that, inferred by a set of pronunciation rules. In languages like English or German, which have a rather complex syllable structure allowing sequences of several consonants, the correct *syllabification* is also a prerequisite for inferring the correct pronunciation.

Finally, the *prosodic structure* of the input text is analyzed, by taking the syntactic structure, POS information, and punctuation into account. Elements of the prosodic structure are phrase breaks, which often entail a short pause as well as a rising pitch contour and lengthening right before the break, and acoustically prominent words and syllables. Prosody also conveys the sentence mode, i.e., the functional type of sentence such as statement and question, as well as the information status of words, such as whether a word provides contextually given information or new information. In English, for instance, questions may have a rising pitch contour at the end of a sentence, whereas statements have a falling contour, and new information will be highlighted with a sentence accent and old information is usually de-accented.

The modules of the text analysis component are based on computational models of various aspects of the grammar of the target language, which is why they are often referred to as computational linguistics or natural language processing modules. Depending on the sophistication of the system and the resources available for the target language, the models can be trained on datasets or embody declarative statements about the properties of the grammar. The output of this component can be characterized as a representation of the input text that is massively enriched by information on many linguistic levels. This representation is the foundation for generating a synthetic speech signal that conveys the linguistic message and its meaning to the listener.

Speech Synthesis

A common distinction of speech synthesis techniques is between *rule-based* and *corpus-based* synthesis methods (see Fig. 5). Rule-based approaches can be further subdivided into acoustic-parametric *formant synthesis* and *articulatory synthesis*. Formant synthesis models the acoustic properties of speech sounds. It is based on the source-filter model (Fant 1960), which means that the acoustic signal is produced by a model that includes control parameters for the voice source (the signal produced by the vibrating vocal folds) and for the vocal tract filter, varying with vocal tract

geometry as a consequence of the position and movements of the speech articulators (in particular the tongue, lips, and lower jaw). The synthesized waveform is modeled by exciting a time-varying filter (representing the vocal tract) by a suitable excitation function (representing the voice source). The intended acoustic shape of the acoustic signal is produced by changing the resonances (*formants*) of the vocal tract filter.

Articulatory synthesis, on the other hand, aims to model the entire process of speech production. The excitation signal, generated by a model of the larynx and the vocal folds, is subsequently filtered and acoustically shaped by a dynamic vocal tract area model. Movements and degrees of freedom of the articulators are also explicitly modeled. The computed resonances can be used to drive a formant synthesizer or some other speech signal generation method. Articulatory synthesis is arguably the scientifically most ambitious synthesis strategy, and it is also maximally flexible with respect to changing the speaker's voice and voice quality, for instance, to convey the speaker's emotional state and even synthesize a singing voice. However, due to the number of approximative solutions, the speech quality produced by articulatory synthesizers tends to be inferior to corpus-based approaches, and the required computations still tend to be ill-suited for real-time applications.

In contrast to rule-based synthesis, corpus-based synthesis makes use of recorded natural human speech. The classical corpus-based approach is *concatenative synthesis* using a fixed set of acoustic units, in particular diphones. Diphones are sequences of two speech sounds, extending from the temporal midpoint of the first sound to the midpoint of the second sound. They preserve the natural, dynamic transitions between speech sounds and concatenate the diphones in locations where the acoustic properties of the speech sounds are relatively stable. Diphone synthesis has gradually been replaced by corpus-based synthesis using acoustic units of variable length, referred to as (nonuniform) *unit selection* synthesis. The key idea of unit selection is to use a large speech database as the acoustic inventory and to select, at system run-time, from this corpus the longest available sequences of phonetic segments that jointly match the speech sound sequence in the target sentence, thereby minimizing the number of concatenations and reducing the need for signal processing. Concatenations tend to produce audible discontinuities in the synthetic waveform, and signal processing tends to degrade the acoustic quality. In an ideal world, the entire target sentence would be available in the recorded corpus and simply played back by the system, effectively rendering natural speech. Given the complexity and combinatorics of language, this ideal situation is very unlikely to arise, but in a database containing several hours of speech, chances are that a target sentence may be synthesized by using just a few units, each being longer than a diphone. Judicious corpus design increases coverage of the language and thereby the probability of finding sequences of long units.

Unit selection synthesis can be quite successful in rendering natural-sounding synthesis. However, it is quite inflexible because a switch to a new speaker or speaking style requires extensive new recordings and database analysis. Another corpus-based technique, known as *statistical-parametric synthesis* (or HMM = hidden Markov model-based synthesis), offers more flexibility. It uses models of speech that can be trained and thus can learn an acoustic mapping function between a

large, existing database and a new database representing a different voice or speaking style. Learning the mapping function requires just a few minutes of target speech. Moreover, the statistical models have an averaging tendency and thus produce smooth transitions between speech units. The disadvantage of this technique lies in its use of a parametric model of the voice source (a *vocoder*) rather than natural excitation signals, yielding a rather buzzy, robot-like synthetic voice quality. However, this is an active research area, and several research groups worldwide are exploring options for improving the voice quality of statistical-parametric synthesis.

Most recently synthesis techniques based on *deep neural nets* have been proposed, which model the entire process of converting text into acoustic speech by means of a complex artificial neural network. The modules of the standard TTS system architecture are substituted by hidden layers of the neural net. This approach has the advantage of learning the mapping from the text representation to acoustic signals without any explicit models of TTS conversion, thereby avoiding the cascading, and potentially reinforcing, detrimental effect of a pipeline of imperfect models of human speech production on the synthetic speech output. From the point of view of speech scientists, this approach is, by and large, a black box, making it difficult for the researcher to inspect sources of error in a diagnostic way.

Resources

Developing a fully fledged TTS synthesizer requires availability of various types of resources, including text and speech databases, tools for linguistic analysis and signal processing, grammars, and dictionaries. A good design principle for TTS systems is to implement a pipeline of language-independent modules that process language-specific linguistic and phonetic data residing in datasets external to the modules. This design has been shown to be particularly efficient in multilingual TTS systems (e.g., Sproat 1998).

But where do the language-specific data come from? Speech, or voice, databases for corpus-based synthesis are constructed by recording natural speech produced by professional speakers, who are sometimes referred to as *voice talents*. The text materials for these recordings are selected from sources representing various text genres, with an optimal coverage of the statistical distributions of the target language in mind. The recordings are subsequently annotated on several linguistic levels and phonetically segmented on the levels of speech sounds, syllables, and words. Other sources of information include dictionaries with information on pronunciation, syllable boundaries, lexical stress, and possibly the morphological structure. The linguistic text analysis component of the TTS system consists of computational models of the grammar of the target language, in particular the syntactic and prosodic structure but also morphological and phonological analyzers to handle out-of-vocabulary words for which no information is available in the dictionary. The models are either based on rules and declarative descriptions written by experts or trained on linguistic data.

In general, there is a strong preference in the field for trainable TTS modules because their performance is usually superior to that of rule-based systems. However, training data for linguistic models, but also for automatic segmentation and annotation of speech databases, are available only for a relatively small number of languages with sufficient resources in terms of databases and formal linguistic descriptions. The training itself relies on language-independent machine learning tools and general-purpose formalisms, such as classifiers, neural nets, HMMs, and finite-state automata. As indicated previously, the latest approach based on deep neural nets learns the complete mapping from text to speech, making grammatical expertise and training data for specific intermediate models potentially obsolete (Shen et al. 2018) and the problem of low-resource languages less pressing.

Applications of TTS Technology

TTS technology can be used in many different situations and application domains, of which we present a selection in this section. TTS technology can be of great communicative help for those who cannot speak, in which case TTS is used as a speech prosthesis. A famous example for such an application is the formant synthesizer that the physicist Stephen Hawking (1942–2018) used. The robotic sound of his synthesizer came to be seen as an important feature of his personality, namely, as his personal voice (personal communication with Lucy Hawking). This aspect is very important when selecting a voice for a speech prosthesis user. For instance, a 10-year-old girl would not like to sound like a 40-year-old woman and definitively not like a man. There is no doubt that a speech synthesizer can be of great help for persons who have lost their voices or articulatory skills. However, handling a synthesizer is rather cumbersome compared to normal speaking, which makes the timing in spoken interaction challenging. An important restriction for the expressivity of the user is that speech synthesizers generate spoken language usually without the possibility of paralinguistic vocalizations such as laughter or sighs and with a restricted range of changing the tone of voice, for instance, a voice with a smile or with anger.

Another field of alternative and augmentative communication is to use TTS systems as reading machines. Individuals who cannot see or have another visual disability can use TTS to convert any electronically available text, e.g., texts on websites, personal texts, or e-mails, into speech. There are of course problems with non-textual content such as pictures, graphs, and visually processed web-based information that is made for seeing people. A further problem is that blind users of synthetic speech would like to get the information at a *fast* speed (similar to seeing people when skimming through printed newspapers). Interestingly, trained blind TTS users are able to listen to, and comprehend, synthetic speech at ultrafast rates, i.e., at speech rates beyond ten syllables per second, which is considered to be the maximum for natural human speech (conversational speech usually shows rates between three and eight syllables per second).

TTS conversion can be used as a device to verify the correct orthography of a written word. Not only blind users can use this function when writing a text but also those who have problems with writing and reading, e.g., dyslexic and illiterate persons. And in second-language learning, TTS can be used to listen to personalized texts containing words with unknown pronunciation.

Situations characterized as "hands busy, eyes busy" and the search for personal information in telephone-based services such as telephone banking are often combined with external information sources. Examples are in-car navigation, the announcement of information in public transportation or airports, or the spoken reproduction of numbers of any kind (telephone numbers, bank account balance, stock prices, station names in public transportation, time information, etc.). Although it often sounds like a playback, it is actually TTS conversion that is used to process this kind of information, because the number of possible utterances can be very large even when the domain and vocabulary are limited. For instance, reading out the string "8328.58 €" requires more than recording the words "eight," "thousand," "three," "hundred," "twenty," "fifty," and "Euro," if it is to sound natural and easy to process for the listener – each *spoken* word, e.g., the three instances of "eight" in the above example, should all have different durations and intonation, depending on the position in the string. Moreover, the word "Euro" should be placed correctly.

Integral parts of spoken dialog systems or voice user interfaces are automatic speech recognition as a technical way to understand the user, speech synthesis to speak to the user, and a dialog manager to access the knowledge sources, to perform the natural language processing and to control the interaction between the user and the system. Dialog systems can operate in closed and in open domains. A ticket-vending machine, e.g., when getting a train ticket from a machine or by phone, belongs to applications with a *closed* domain. The vocabulary is very limited and the possibility of sentences to be produced is limited, too. The examples from information search also pertain to closed domains. A personal assistant like those used in smartphones or other devices in the home, often in the design of a loudspeaker, acts in an *open* domain. In principle, there are no limitations regarding the vocabulary and the way users formulate their questions and commands to the personal assistant.

Further applications include narratives in computer games and automated audio books. Here, the artificial voices need additional individual distinctions of personality beyond a pure TTS conversion. Other everyday applications of TTS are, for instance, translation devices on the web which provide, among other information, the pronunciation of words and names not only in phonetic script but also in the form of an audio file. This special feature can be considered as a great improvement over printed dictionaries.

Evaluation of Synthetic Speech

The fact that from an engineering perspective it is possible to synthesize humanlike speech does not mean that human users of synthetic speech find the spoken output of a speech synthesizer useful and easy to handle. The evaluation of synthetic speech

quality and its appropriateness for the needs of the user remains a challenging task. For instance, a blind user with several years of experience with TTS has substantially different ideas of intelligibility and naturalness than a novice user listening to the spoken output of short machine-translated messages. Furthermore, limitations in expressivity may have an impact on how people enjoy spoken human-machine interaction.

Typically, two dimensions are central for the listener: intelligibility and naturalness. It is important to note that intelligibility comes first and naturalness second, since a highly naturally sounding message that nobody can understand is useless. Intelligibility can be impeded by many different factors such as an incorrectly placed accent, a tempo that is too fast or poor signal quality. Further questions that can be asked when evaluating synthetic speech are how attractive computerized voices are and how much fun it is to use synthetic speech. Quite often these aspects are not rated very high.

A crucial point for evaluation is who is the evaluator. End users may have a completely different opinion than developers. For the latter group, it has been reported informally that after half an hour of working with synthetic speech, a developer or researcher will find the quality very good. The users at the other end can show a great diversity when assessing the output of TTS systems. A blind user with many years of experience will behave differently compared to an unexperienced user who went blind at an old age. Likewise, a listener whose mother tongue is identical with the synthesized language will have fewer problems in understanding than a second-language listener; this tends to be true even for speakers with an advanced proficiency level.

The evaluation itself can take place at different levels which are often independent of each other. Of course, the quality of the voice plays a vital role. Does the artificial voice sound robotic, and how pleasant is it to listen to this voice? Then, matters of pronunciation come into play. How are words pronounced that are out of vocabulary and that are morphologically complex, and how are abbreviations treated? Questions of prosody play a role on the word and sentence level. Is the word stress correct, are the sentence accents acceptable, is the phrasing adequate, are rhythm and tempo pleasant, are punctuation and graphical highlighting in the text converted correctly? On the sentence and text level, the discourse structure is a key factor. Does the sentence accent structure reflect the correct focus, is new information prosodically marked, can the listeners perceive a contrast when present? Finally, the question may be asked whether the selected speech style is appropriate for the purpose. An e-mail with a happy content read out in the style of listing stock options does not seem to be adequate, just as a sad message expressed with a lively and enthusiastic voice.

One good example for a regularly performed evaluation of speech synthesis is the "Blizzard challenge" (organized by the Special Interest Group on Speech Synthesis of the International Speech Communication Association, ISCA). It is designed as a competition between different TTS developer teams who have access to the same English voice database. The task for every developer is to generate the same speech material, mostly single sentences. In the worldwide evaluation, experts and laymen take part in an online test that takes about 30 min and is structured into different

parts. In parts with longer speech sections (up to 30 s), the listeners rate with a slider the following dimensions: overall impression (bad, excellent), pleasantness (very unpleasant, very pleasant), pauses (confusing, appropriate), stress (unnatural, natural), intonation (fit, do not fit the sentence type), emotion (no expression of emotion, authentic expression), and listening effort (very exhausting, very easy). In another part single sentences are rated on a five-point scale according to their naturalness. Yet another task is to score the similarity of two synthetic voices. To measure intelligibility semantically unpredictable sentences are used, and listeners are asked to type as many words as they understood.

It becomes clear that evaluation is a substantial part of TTS system development that needs considerable human resources in the form of raters. This fact makes evaluation a time-consuming and expensive task. Therefore, the evaluation is often made via crowd sourcing where many listeners perform easy tasks for little money. Although these subjective evaluations come close to the final purpose of speech synthesis, namely, the listener, TTS developers are for various reasons also interested in objective evaluations. The ultimate goal is to develop standard metrics of synthesis quality that allow a prediction of subjective assessment by human listeners.

Further Modes in Artificial Speech Communication

As outlined previously, *articulatory synthesis* is the most complete and explicit simulation of the principles of human speech production. The simulation also pertains to the radiation of an acoustic pressure wave at the nostrils and lips as a function of the vocal tract geometry, which is controlled by the position and movement of the articulators and excited by one or more sound sources, viz., an excitation signal generated by the vocal folds and optionally turbulent noise produced at constrictions in the vocal tract. An alternative method of generating speech signals based on articulatory information uses methods that are well-established in automatic speech recognition and in statistical-parametric speech synthesis. In this approach, known as the *silent speech interface* (Denby et al. 2010), the mapping between articulatory gestures and patterns in the acoustic speech signal is learned from natural acoustic and articulatory data. At synthesis run-time, the movement of articulators is tracked by electromagnetic sensors, and a synthetic speech signal is generated based on the statistical models. There are at least two prominent application scenarios for this approach. First, a silent speech interface can be useful in a situation in which confidential information (such as a password) is conveyed to a technical system in a public space. In this case the silent speech interface would transmit a transcript of the spoken utterance without actually generating overt speech. A variant of this scenario is the transmission of spoken language to an interlocutor across a phone connection when the content is confidential or overt speech would be masked by ambient noise. In this variant, the utterance covertly produced by the speaker is synthesized at the remote end of the connection. Second, a silent speech interface can synthesize speech for speakers who are unable to produce the natural excitation signal required for audible speech but are still able to articulate. Finally,

direct communication pathways from the brain to a speech synthesizer are being explored in the context of brain-computer interfaces, based on either wired connections or, preferably, noninvasive imaging techniques such as functional near-infrared spectroscopy. Such *brain-to-speech* systems convert brain waves to synthetic speech with no need for overt articulatory gestures (Herff et al. 2015).

TTS synthesis and synthesis methods that take other types of representations of the intended linguistic message as input, such as articulatory gestures or brain waves, are essentially unimodal, as they typically generate a synthetic speech signal. However, natural communication is multimodal. Interlocutors engaged in a face-to-face conversation produce and perceive speech-related information at least in two modalities, the acoustic and the visual channel, and in several modes, such as the speech mode and speech-accompanying gestures and facial and other body movements. It is a well-established fact that multimodal speech is not only more natural but indeed also more intelligible than speech transmitted by the acoustic channel alone. *Audio-visual speech synthesis* is an active research area using different approaches for modeling the two modalities and their synchronization. Ultimately, the goal is to generate both the acoustic speech signal and the accompanying gestures jointly, based on a unified model of multimodal speech production. The SmartKom system (Wahlster 2006) is an early example of a dialog system that combines multimodal information in both directions, viz., speech recognition and speech synthesis. In this experimental system, the human interlocutor may use natural combinations of spoken language, conventional gestures such as pointing to objects in the environment or on a map, and facial expressions to interact with the system. Conversely, the system generates synchronized multimodal output, exploiting the synergy between spoken language and other means of referring to objects. For instance, it is natural to use an emphasized expression like "this movie theater" or just "here" while pointing to an object on a map, rather than specifying the identity of the object explicitly, and redundantly, in the spoken utterance. The SmartKom project also included information inferred from speech about the speaker's emotional state but did not attempt to synthesize affective speech. The latter is another hot topic in speech synthesis research but beyond the scope of this overview.

SmartKom used a cartoon persona as a means of conveying speech-accompanying gestures. More generally, talking heads, avatars, and other animated personas are frequently used in embodied speech synthesis and dialog systems. Speech synthesis is also an essential capability of robots interacting with humans in a naturalistic way, a field that presumably will gain more and more relevance in the years to come.

TTS Across the World

Although there has been much progress in speech synthesis quality in the last decade, often achieved by processing enormous amounts of data, TTS today is available only for relatively few languages and dialects. The main reason for this unsatisfying state is that there are thousands of languages with no or few linguistic

and speech resources. For instance, when considering the 31 official and semiofficial languages of the European Union, the speech and text resources for speech and language technology for 21 languages are fragmentary, weak, or simply not existing, nine languages have a moderate support (Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish), and only one language (English) is considered well supported (Rehm and Uszkoreit 2013). In addition, there are regional and minority languages which all are threatened by digital language extinction. Presumably, many European languages are among the better-resourced languages in the world, although several languages with a smaller number of speakers can be considered as under-researched and under-resourced. Globally, many under-resourced languages are non-scripted, i.e., they do not have a writing system, which makes the conversion of text to speech impossible. For languages with nonalphabetical writing systems, a translation or a systematic transcription is often required for further TTS conversion. This is either done automatically or the user types the text in a romanized form, e.g., Pinyin for Mandarin Chinese.

Since TTS conversion is bound to text, sign languages cannot be directly converted into speech, and a translation step is required first. Converting text to a specific sign language, such as American Sign Language, British Sign Language, or German Sign Language, is a special form of language synthesis. It is independent of acoustics, but properties attributed to spoken language such as prosody are also an important feature of signed language that must not be neglected in sign language synthesis.

Although there are several thousands of languages and language varieties, TTS research and commercial solutions only exist for relatively few languages. It is no surprise that English, and in particular the American and British varieties, dominate research and development, complemented by maybe a dozen other languages. Commercially, TTS is offered for a slightly larger number of languages, but still thousands of languages are not considered. Thus, a major task will be to find or create linguistic resources and make them available in an adequate way for many languages and language varieties. This also includes appropriate ways of distribution and accessibility. Creating resources for under-resourced languages is an active research area with its own series of conferences (see the International Speech Communication Association Special Interest Group on Under-resourced Languages).

Conclusions

It is evident that TTS technology is now available in everyday applications, not only for technical specialists and for special user groups but for the general public, too. It is a standard feature on mobile communication devices, in home electronics and appliances, in cars and public transportation, and in commercial dialog and information systems. In this sense, it is often characterized as a mature technology.

However, TTS is not a solved problem from a scientific perspective by any means. All currently available synthesis methods generate speech output that deviates in characteristic ways from natural speech. For instance, concatenative synthesis produces audible discontinuities in the artificial waveforms that cannot be produced

by a human vocal tract, and our auditory system is extremely sensitive to these oddities. Statistical-parametric synthesis, on the other hand, succeeds in avoiding discontinuities and producing smooth acoustic trajectories (some say, too smooth to be natural) but has a robot-like voice quality. From a scientific point of view, regarding TTS as a model of human speech production, these imperfections suggest that there are still quite a few properties of natural speech, and how it is produced, that are not yet properly understood.

A number of other challenges remain, too. One is the evaluation of the quality of synthetic speech and its adequacy for the needs of the user. Developing objective quality measures that successfully reproduce, or predict, the assessment of synthesis quality by human listeners remains an elusive goal. Part of the problem is that properties of speech such as intelligibility, naturalness, and pleasantness must be evaluated in many linguistic and phonetic dimensions. A second challenge is to develop synthesis methods that are flexible enough to generate different speaking styles, such as neutral reading of newspaper text or engaging, expressive conversational speech, or emotionally colored speech. Switching between varieties of languages and dialects is another capability that is still by and large lacking in TTS systems.

Especially in the light of trainable TTS components, the biggest challenge is the lack of text and speech data and formal linguistic descriptions for the vast majority of languages. The availability of intuitive spoken language-based interfaces to technology is a key factor for access to information and knowledge in societies around the world.

References

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication, 52*(4), 270–287.

Dudley, H. (1940). The carrier nature of speech. *The Bell System Technical Journal, 19*(4), 495–515.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience, 9*, 217. https://doi.org/10.3389/fnins.2015.00217. Accessed 01 Aug 2018.

Rehm, G., & Uszkoreit, H. (Eds.). (2013). *The META-NET strategic research agenda for multilingual Europe 2020*. Heidelberg: Springer.

Shen, J., Pang, R., Weiss, R. J., Schuster M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of IEEE international conference on acoustics, speech and signal processing*, Calgary, paper #3782.

Sproat, R. (Ed.). (1998). *Multilingual text-to-speech synthesis – the Bell Labs approach*. Dordrecht: Kluwer.

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

von Kempelen, W. (2017). *Mechanismus der menschlichen Sprache – The Mechanism of Human Speech.* Kommentierte Transliteration & Übertragung ins Englische – Commented Transliteration & Translation into English by Fabian Brackhane, Richard Sproat & Jürgen Trouvain (Eds.). Dresden: TUDpress. Also available online http://www.coli.uni-saarland.de/~trouvain/kempelen.html

Wahlster, W. (Ed.). (2006). *SmartKom: Foundations of multimodal dialogue systems*. Berlin: Springer.