RESEARCH ARTICLE OPEN ACCESS

Tutorial - Speech Synthesis System

Sangramsing Kayte*,Siddharth Dabhade**
Department of Computer Science and Information Technology, Dr.Babasaheb Ambedkar Marathwada University, Aurangabad.
Department of Computer Science and Information Technology, Dr.Babasaheb Ambedkar Marathwada University, Aurangabad

ABSTRACT
Speech synthesis we can, in theory, mean any kind of synthetization of speech. For example, it can be the process in which a speech decoder generates the speech signal based on the parameters it has received through the transmission line, or it can be a procedure performed by a computer to estimate some kind of a presentation of the speech signal given a text input. Since there is a special course about the codecs (Puheen koodaus, Speech Coding), this chapter will concentrate on text-to-speech synthesis, or shortly TTS, which will be often referred to as speech synthesis to simplify the notation. Anyway, it is good to keep in mind that irrespective of what kind of synthesis we are dealing with, there are similar criteria in regard to the speech quality. We will return to this topic after a brief TTS motivation, and the rest of this chapter will be dedicated to the implementation point of view in TTS systems. Text-to-speech synthesis is a research field that has received a lot of attention and resources during the last couple of decades – for excellent reasons. One of the most interesting ideas (rather futuristic, though) is the fact that a workable TTS system, combined with a workable speech recognition device, would actually be an extremely efficient method for speech coding). It would provide incomparable compression ratio and flexible possibilities to choose the type of speech (e.g., breathless or hoarse), the fundamental frequency along with its range, the rhythm of speech, and several other effects. Furthermore, if the content of a message needs to be changed, it is much easier to retype the text than to record the signal again. Unfortunately this kind of a system does not yet exist for large vocabularies. Of course there are also numerous speech synthesis applications that are closer to being available than the one discussed above. For instance, a telephone inquiry system where the information is frequently updated, can use TTS to deliver answers to the customers. Speech synthesizers are also important to the visually impaired and to those who have lost their ability to speak. Several other examples can be found in everyday life, such as listening to the messages and news instead of reading them, and using hands-free functions through a voice interface in a car, and so on.
Keywords **-** Speech Coding, speaker, single, Phonetic analysis.

I. INTRODUCTION

The most commonly used criteria for high-quality speech are intelligibility, naturalness and pleasantness. Since these are multidimensional factors that depend on each other, the comprehensive high quality is formed by the interaction of numerous details. Thus, the elimination of background noise, musical noise, mumbling, and the various pops and cracks, does not result in the ultimate quality, but the speech should also be made rich in nuances and it should carry information about the personality of the speaker. Moreover, it would be advisable to include in the speech some features that describe the emotional state of the speaker because this improves the naturalness and makes the speech livelier. The quality of speech synthesizers can also be examined apart from speech quality as such, in which case we are interested in, for instance, the following facts: How much the operation of the synthesizer depends on the surrounding software or the operating system? How good is the performance of the synthesizer in the particular application for which it has been designed? How much memory does it require? Given the continuously increasing processing capacity, more and more complicated speech synthesizers can be used even on personal computers.As mentioned in the context of speech enhancement, it is a more or less vague task to measure the quality of speech. No single method can be specified that would give absolutely correct and comprehensive results. A sufficient amount of information about the speech quality produced by a certain system can often be obtained by performing several different tests and combining their results. The problem is that there is a wide range of test methods available, and the uncertainty of measuring can be rather high, and therefore it can be problematic to compare the test results published by different directions. Besides, arranging the tests – especially listening tests – is usually very laborious and expensive.

If we ignore the costs and other problems, it is naturally advisable to examine the quality of

synthesized speech from several angles. This makes it possible to draw conclusions about the performance of the different modules of the synthesizer, such as the linguistic or prosodic analysis (more about these in a while). The intelligibility of speech can be estimated by, e.g., testing how well the listeners can discriminate different consonants from each other in synthesized speech. After all, it is more difficult to produce consonants than vowels by synthesis methods. To avoid context dependency, the utterances used in these listening test are usually very short, like syllables, or nonsense words. When studying the longer-term understandability of speech, the test utterances are typically sentences so that the content of the utterances can be understood despite some errors in certain phonemes. The difficult part is to evaluate how accurately the prosodic features have been attached by the synthesizer. This is done by listening tests that somehow explore the opinions of the listeners in regard to the intended and realized prosodic patterns in the test utterance. In contrast, there are quite well-established and frequently used methods for measuring the general speech quality, i.e., the amount of audible distortion, mumbling, rasp and other undesired effects. The different test methods will not be described here but information about the most popular objective measurements and listening tests can be found practically in any text book related to speech signal processing.

In the present-day speech synthesizers, there are still several limitations concerning speech naturalness and personality. However, intelligibility has already reached a high level, which makes it possible to use speech synthesizers in certain applications. Besides, intelligibility can be improved by presenting facial animations together with the synthesized speech. This so called audiovisual speech synthesis is actually one of the latest trends in the synthesis research (Lemmetty, 1999).

II. IMPLEMENTATION OF TTS

The process of transforming text into speech contains coarsely two phases: first the text goes through analysis and then the resulting information is used to generate the speech signal. In the block diagram shown in Figure 1, the former phase actually contains not only text analysis but also phonetic analysis in which the graphemes are converted into phonemes. The generation of the speech signal can also be divided into two sub-phases: the search of speech segments from a database, or the creation of these segments, and the implementation of the prosodic features. These phases will be further discussed in the following.

Text analysis is all about transforming the input text into a 'speakable' form. At the minimum, this contains the normalization of the text so that numbers and symbols become words, abbreviations are replaced by their corresponding whole words or phrases, and so on. This process typically employs a large set of rules that try to take some language-dependent and context-dependent factors into account. The most challenging task in the text analysis block is the linguistic analysis which means syntactic and semantic analysis and aims at understanding the content of the text. Of course, a computer cannot understand the text as humans do, but statistical methods are used to find the most probable meaning of

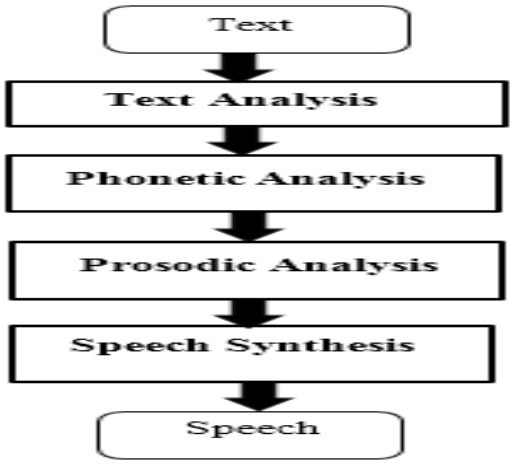

Figure 1: Block diagram of text-to-speech synthesis.

The utterances. This is important because the pronunciation of a word may depend on its meaning and on the context (for instance, the word record is pronounced in different ways depending on whether it is a verb or a noun). Finally, the text analysis block is supposed to provide prosodic

information to the subsequent stages. It can, for example, signify the positions of pauses based on the punctuation marks, and distinguish interrogative clauses from statements so that the intonation can be adjusted accordingly. Phonetic analysis converts the ortographical symbols into phonological ones using a phonetic alphabet. We have already seen on this course IPA, the alphabet of the International Phonetic Association. IPA contains not only phoneme symbols but also diacritical marks and other symbols related to pronunciation. Since the IPA symbols are rather complicated and there are several symbols that cannot be found in typewriters, other phonetic alphabets have also been developed. They are better compatible with computers and often based on ASCII characters. Examples of such alphabets are SAMPA (Speech Assessment Methods – Phonetic Alphabet), Worldbet and Arpabet. However, there is no generally accepted, common phonetic alphabet and therefore separate speech synthesizers often use their own special alphabets (Lemmetty, 1999). The degree of challenge in phonetic analysis is strongly language dependent – Finnish is actually one of the easiest languages in this respect because the pronunciation is not so different from the written form of the utterance. Prosody is a concept that contains the rhythm of speech, stress patterns and intonation. The attachment of certain prosodic features to synthetic speech employs a set of rules that are based on the prosodic analysis of natural speech. Prosody plays a very important role in the understandability of speech and furthermore, prosodic features carry lots of information about the speaker, for instance his or her emotional state, and even the social background. In practice, generating natural sounding prosody in large vocabulary speech synthesis is still a distant goal because the modeling of prosody is such a problematic task. Some hierarchical rules have been developed to control the timing and fundamental frequency, and this has made the flow of speech in synthesis systems somewhat more natural sounding. Speech synthesis block finally generates the speech signal. This can be done either based on a parametric representation, in which case phoneme realizations are produced by machine, or by selecting speech units from a database. In the latter method, a sophisticated search process is performed in order to find the appropriate phoneme, diphone, triphone, or other unit at each time. Whichever method is chosen, the resulting short units of speech are joined together to produce the final speech signal. One of the biggest challenges in the synthesis stage is actually to make sure that the units connect to each other in a continuous way so that the amount of audible distortion is minimized.

A. ***Formant Synthesis***

This is the oldest method for speech synthesis, and it dominated the synthesis implementations for a long time. Nowadays the concatenative synthesis is also a very typical approach. Formant synthesis is based on the well-known source-filter model which means that the idea is to generate periodic and non-periodic source signals and to feed them through a resonator circuit – or a filter – that models the vocal tract. The principles are thus very simple, which makes formant synthesis flexible and relatively easy to implement. In contrast to the methods described below, formant synthesis can be used to produce any sounds. On the other hand, the simplifications made in the modeling of the source signal and vocal tract inevitably lead to somewhat unnatural sounding result.

In a rudely simplified implementation, the source signal can be an impulse train or a sawtooth wave, together with a random noise component. To improve the speech quality and to gain better control of the signal, it is naturally advisable to use as accurate model as possible. Typically the adjustable parameters include at least the fundamental frequency, the relative intensities of the voiced and unvoiced source signals, and the degree of voicing. The vocal tract model usually describes each formant by a pair of filter poles so that both the frequency and the bandwidth of the formant can be determined. To make ingelligible speech, at least three lowest formants should be taken into account, but including more formants usually improves the speech quality. The parameters controlling the frequency response of the vocal tract filter – and those controlling the source signal – are updated at each phoneme. The vocal tract model can be implemented by connecting the resonators either in cascade or parallel form. Both have their own advantages and shortcomings but they will not be discussed here. In addition to the resonators that model the formants, the synthesizer can contain filters that model the shape of the glottal waveform and the lip radiation, and also an anti-resonator to better model the nasalized sounds.

B. ***Concatenative Synthesis***

This is the so called cut and paste synthesis in which short segments of speech are selected from a pre-recorded database and joined one after another to produce the desired utterances. In theory, the use of real speech as the basis of synthetic speech brings about the potential for very high quality, but in practice there are serious limitations, mainly due to the memory capacity required by such a system. The longer the selected units are, the fewer problematic concatenation points will occur in the synthetic speech, but at the same time the memory

requirements increase. Another limitation in concatenative synthesis is the strong dependency of the output speech on the chosen database. For example, the personality or the affective tone of the speech is hardly controllable. Despite the somewhat featureless nature, concatenative synthesis is well suited for certain limited applications.

What is the length of the selected units then? The most common choices are phonemes and diphones because they are short enough to attain sufficient flexibility and to keep the memory requirements reasonable. Using longer units, such as syllables or words, is impossible or impractical for several reasons. The use of diphones in the concatenation provides rather good possibilities to take account of coarticulation because a diphone contains the transition from one phoneme to another and the latter half of the first phoneme and the former half of the latter phoneme. Consequently, the concatenation points will be located at the center of each phoneme, and since this is usually the most steady part of the phoneme, the amount of distortion at the boundaries can be expected to be minimized. While the sufficient number of different phonemes in a database is typically around 40 – 50, the corresponding number of diphones is from 1500 to 2000 but a synthesizer with a database of this size is generally implementable (Lemmetty, 1999). On the other hand, the use of phonemes is the most flexible way of generating various utterances, at least if we ignore the fact that certain phonemes (e.g., plosives) are fairly impossible to separate from a speech signal to their own segments.

In both phoneme and diphone concatenation, the greatest challenge is the continuity. To avoid audible distortions caused by the differences between successive segments, at least the fundamental frequency and the intensity of the segments must be controllable. The creation of natural prosody in synthetic speech is impossible with the present-day methods but some promising methods for getting rid of the discontinuities have naturally been developed. Finally, concatenative speech synthesis is afflicted by the troublesome process of creating the database from which the units will be selected. Each phoneme, together with all of the needed allophones, must be included in the recording, and then all of the needed units must be segmented and labeled to enable the search from the database. Some of these operations can be automatized to certain extent.

C. ***Articulatory Synthesis***

Compared with the other synthesis methods presented in this chapter, articulatory synthesis is by far the most complicated in regard to the model structure and computational burden. The idea in articulatory synthesis is to model the human speech production mechanisms as perfectly as possible. The implementation of such a system is very difficult and therefore it is not widely in use yet. Experiments with articulatory synthesis systems have not been as successful as with other synthesis systems but in theory it has the best potential for high-quality synthetic speech. For example, it is impossible to use articulatory synthesis for producing sounds that humans cannot produce (due to human physiology). In other synthesis methods it is possible to produce such sounds, and the problem is that these sounds are usually perceived as undesired side effects. The articulatory model also enables more accurate transient sounds than other synthesis techniques.

Articulatory synthesis systems contain physical models of both the human vocal tract and the physiology of the vocal cords. It is common to use a set of area functions to model the variation of the cross-sectional area of the vocal tract between the larynx and the lips. The principle is thus similar to the one that has been seen within the acoustic tube model. The articulatory model involves a large number of control parameters that are used for the very detailed adjustment of the position of lips and tongue, the lung pressure, the tension of vocal cords, and so on. The data that is used as the basis of the modeling is usually obtained throug the X-ray analysis of natural speech (Lemmetty, 1999). As expected, such analysis is also very troublesome.

III. FURTHER INFORMATION

There are several excellent summaries about the commercial and non-commercial speech synthesizers and their history, and the purpose of this section is not to repeat them in detail. An interested reader can find more detailed information from literature (a couple of references will be named at the end), or from the internet. A very short recap was, however, considered useful at this point. The first corner-stone of the history of speech synthesizers is generally thougth to be Voder (Voice Operating Demonstrator), the first electrical speech synthesizer, published in 1939 by Homer Dudley who worked for Bell Laboratories. The operation of Voder was controlled by hand. The user chose either a noise-like or a periodic source signal and fed it into a set of band-pass filters whose amplitude responses were controlled by fingers. The fundamental frequency was adjusted by a foot pedal. (There were not many people who had the skills to play this machine.) Voder has been considered to be the first indication that speech can be produced artificially, and it motivated several research groups to invest energy in speech synthesis. The earliest versions of formant and articulatory synthesis systems were published in the 1950's. After the introduction of digital computers, the first full TTS system for the English language came out in Japan in 1968. It was

based on an articulatory model and it has been stated to produce rather intelligible, but monotonic, speech (Huan, Acero, Hon, 2001). Formant synthesis (for example, in Klattalk) dominated the field of speech synthesis for quite a long time, but the introduction of PSOLA (Pitch-Synchronous Overlap-Add) in 1985 considerably facilitated the research and development of concatenative synthesis systems. The idea in PSOLA is to extract speech frames pitch-synchronously, i.e., the center of each frame is located at the pitch pulse position (the highest peak within a pitch period). At the synthesis stage these frames are partly overlapped and summed so that the desired time- and pitch-scale are realized. This way the prosodic features of speech can be adjusted independently from each other. For example, if the speech rate is to be lowered while keeping the pitch untouched, certain analysis frames are duplicated to fill the new, lengthened time axis, but the distance between adjacent synthesis frames is the same as that of the analysis frames, which results in an unchanged pitch frequency. Correspondingly, pitch can be raised by setting the synthesis frames closer to each other than the analysis frames were. To produce opposite effects, it is likewise possible to eliminate some frames instead of duplicating them. As can be anticipated, PSOLA is best applicable to voiced speech in which the pitch period can be determined. As a matter of fact, PSOLA is very sensitive to errors in the pitch estimate, which often causes problems in practice. On the other hand, PSOLA is extremely simple and therefore it is computationally very feasible. In its simplest form, the time-domain PSOLA, the frames are not even modified between the analysis and synthesis stages, but instead, all effects are produced by controlling the number of frames and the distances between them. Matlab implementations of PSOLA can be found, for instance, from the web page of the Digital Audio Effects book (see references)

http://www.unibw-hamburg.de/EWEB/ANT/dafx2002/DAFX_Book_Page/matlab.html where Chapter 7 contains PSOLA. Different variations of the basic PSOLA are commonly used in concatenative speech synthesis. Some experiments have also been made to apply the 'harmonic plus noise model' in speech synthesis. In this model, the speech signal is presented as the sum of harmonically related sinusoidals and a noise component. Actually this method is somewhat better applicable to singing voice synthesis, though. Other examples of the methods tested within speech synthesis are Hidden Markov Models (HMM) and neural networks. Both have been successfully applied to speech recognition but they are also believed to have some potential for speech synthesis purposes.

Finally, it must be emphasized that speech synthesis is related to numerous research fields. Even though the signal processing methods alone form a wide research area, several other disciplines are also needed, and it is necessary to strengthen the co-operation between disciplines in order to achieve even better speech synthesizers. Some of the most essential research areas are phonetics, grammar, lexicology, semantics, pragmatics, data retrieval, software engineering, and signal processing.

REFERENCES

[1] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.
[2] S. Lemmetty, Review of Speech Synthesis Technology, Master's Thesis, Helsinki University of Technology, 1999.
[3] U. Zölzer (editor), DAFX - Digital Audio Effects, John Wiley & Sons, Ltd, 2002.