2018 Eighth International Conference on Instrumentation and Measurement, Computer, Communication and Control

An overview of speech synthesis technology

Yin Zhigang
Institute of Linguistics,
Chinese Academy of Social Sciences
Beijing, China,
yinzhg@cass.org.cn

***Abstract*—Speech is the most natural and convenient approach of communication and speech synthesis technology is a kind of import application in Human-machine interaction system. This paper gives a comprehensive overview of Text-to-Speech (TTS) synthesis technology. The two basic parts of speech synthesis technology are natural language processing (NLP) and digital signal processing (DSP). To the part of NLP, some important steps are pre-processing, morphological analysis, contextual analysis, syntactic-prosodic analysis, phonetization and prosody generation. To the part of DSP, two types of synthesis methods are rule-driven methods and data-driven methods. Some important synthesis approaches of DSP such as articulatory synthesis, formant synthesis, concatenative synthesis, unit selection synthesis, HMM synthesis and DNN synthesis are introduced. Finally, these approaches of speech synthesis are compared briefly. The technical trends of TTS and some hot spots of its applications in the future are discussed.**

Keywords—speech synthesis; text-to-speech; natural languages processing; digital signal processing; rule-driven; data-driven.

I. INTRODUCTION

Speech is the most natural and convenient approach of communication between humans and machines. Two basic parts of human-machine interaction system are speech recognition and speech synthesis. Speech recognition is the input part for human-machine interaction system, and it is called ASR (Automatic Speech Recognition) technology too. The output part is speech synthesis which is also called Text-to-Speech (TTS) technology because this is a program to automatically generate speech from text.

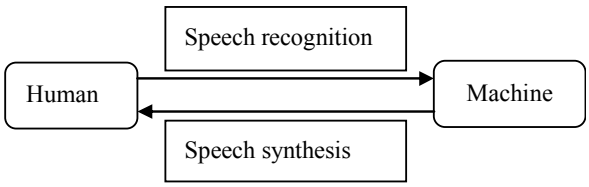

Fig. 1. The Human-Machine Interaction System

In this paper, we will give a brief but comprehensive overview of speech synthesis techniques, and discuss the trends of technology and applications.

II. FROM THE TEXT TO PHONETIC REPRESENTATION

There are two phases in the procedure of speech synthesis. One of them is text-analysis phase, where the input text is transcribed into phonetic representation. This process is also called natural language processing (NLP). The other phase is the generation of speech sounds, where the speech sounds are produced from the phonetic representations. This process is also called digital signal processing (DSP). Fig.2 shows the two phases of Text-to-Speech synthesis.

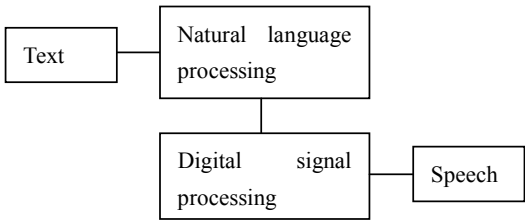

Fig. 2. Two phases of speech synthesis

In the phase of text-analysis, the text will be analyzed as linguistic units such as sentences, phrases, words and then be transcribed into phonetic representation. Some important steps in this phase are pre-processing, morphological analysis, contextual analysis, syntactic analysis, phonetization and prosody generation. Fig. 3 shows these steps in natural language processing.

The pre-processing block will correct errors or irregularities of the input text. It converts abbreviations, numbers and acronyms into full text and then breaks the text into sentences. Usually the sentences' boundaries could be indicated by punctuations. But for those texts without clear boundaries' indicators, some classifier techniques should be used to determine the boundaries.

The morphological analysis block and contextual analysis block usually work together. They divide the sentence into a sequence of words and categorize each word into possible part-of-speech on the basis of the word's spelling. The first step is also called word segmentation or word tokenization. For some languages such as English, the whitespace is a clear mark of word boundaries. But for other languages such as Chinese, there are no spaces between words as delimiters. The word segmentation of these languages would rely on the contextual analysis. The contextual analysis block could also streamline the list of possible parts of speech of words in sentences, by considering the parts of speech of neighboring words.

The syntactic analysis block parses the syntactic structure of a sentence according to the information of word segmentation and word classes. An important aim of this step

978-1-5386-8246-3/18/$31.00 ©2018 IEEE
DOI 10.1109/IMCCC.2018.00116

is to predict the prosody of the input sentence so this step is also called syntactic-prosodic analysis.

The phonetization block generates the sequence of phonetic symbols for each word. A pronunciation lexicon should be constructed first and the phonetic symbols of each word should be produced according to the lexicon. For those words such as names and special characters that could not be covered by the lexicon, the special algorithm which could generate the phone sequence according to the morphology factors of the words is necessary. As a result, phonetization block is an integration of dictionary-based methods and rule-based methods.

The last step is prosody generation. It generates the prosodic features which include pitch curve, duration and pause information, stress and rhythm features. The prosody performance is considered to have a significant impact on the naturalness degree of TTS. In fact, it also affects the intelligibility degree of TTS because the prosodic features could help listeners to segment sentences into chunks comprising of groups of phrases, words and syllables.

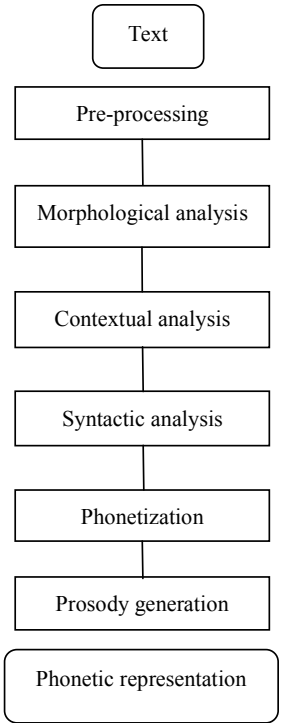

Fig. 3. The module of Natural language processing

After the above steps, an input text is transcribed into corresponding phonetic representations. The next task is to generate the speech sounds from these phonetic representations.

III. From Phonetic Representation to Speech

In the phase of generation of speech sounds, there are two types of synthesis methods: rule-driven methods and data-driven methods. The principle of rule-driven methods is to simulate the speech voice according to the rules deduced from articulation process or acoustic process. The principle of data-driven methods is to generate the speech by recorded speech data or the statistical parameters got from speech data.

There are several approaches in each of the two types of TTS methods. The two approaches of Rule-driven synthesis methods are articulatory synthesis and formant synthesis. The main approaches of Data-driven synthesis methods are concatanative synthesis, unit selection synthesis, HMM synthesis and DNN synthesis.

Fig. 4 shows these synthesis methods and approaches.

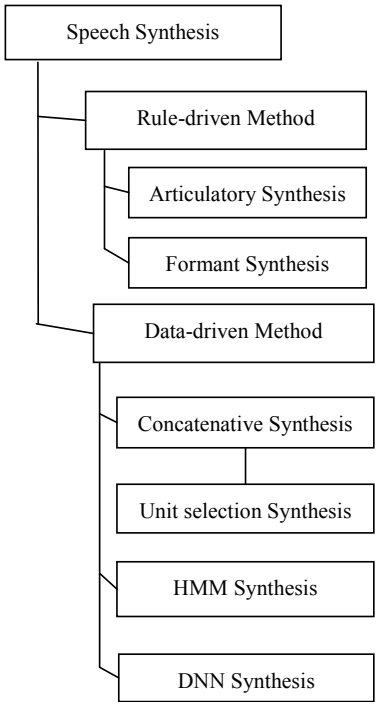

Fig. 4. The methods and approaches of Speech synthesis

The two important indexes to evaluate the qualities of speech synthesis systems are the intelligibility and the naturalness. In general, Data-driven methods are superior to Rule-driven methods in both two respects because they use real speech data. However, Rule-driven methods have greater advantages in data capacity and synthesis flexibility.

Each approach is described in detail below.

A. Articulatory Synthesis

Articulatory Synthesis is a direct simulation of the physical process of human pronunciation. Usually, a series of rules to manipulate the articulatory model should be developed first. Its advantage is very flexible. Developer can adjust the

parameters of the model precisely and change the synthesized sounds easily. But its main disadvantage is the sound quality is not good because the human's pronunciation process is very difficult to be simulated accurately. The control parameters of the model include lip aperture, lip protrusion, tongue tip position, tongue tip height, tongue position and tongue height, etc [1]. Not only acquiring data for articulatory model but also designing and controlling the model with the parameters are difficult. And the other difficulty is to control the high level parameters such as prosodic features.

B. *Formant Synthesis*

Format synthesis is a simulation of acoustic process. Formant is the resonance frequency area of the vocal tract. The formants constitute the main frequencies that make sounds (especially vowels) distinct. Commonly each vowel has 4-5 visible formants in the spectrum.

Formant synthesis models the pole frequencies of speeches. It is based on the source-filter model of speech production. In this model, speech sound is produced by two kinds of sound sources: the source of voiced sound (such as vowels) is a periodic signal with a fundamental frequency, and the source of unvoiced sounds (most of consonants) is a random noise generator. Then, the voiced and unvoiced sounds generated from sound sources are modified by the vocal tract model and produced by the radiation model (amplifier).

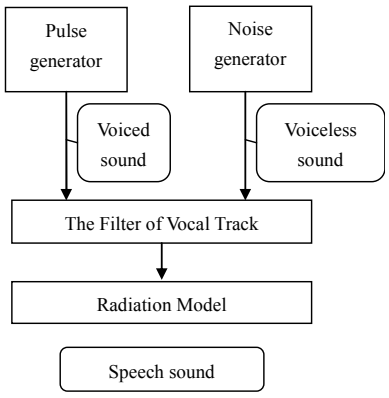

Fig. 5. Source-filter model in pronunciation

In 1980 Dennis Klatt developed one of the most famous formant synthesizers: Klatt synthesizer. It was controlled by 39 parameters which were updated every 5 ms [2][3]. Format synthesis also has the advantages of Rule-driven methods, small and flexible. The sound quality of formant synthesis is better than that is produced from articulatory synthesizer but the produced speech is still far from natural.

In order to get the natural speech, conctenative synthesis was proposed at the end of 20th century.

C. *Concatenative Synthesis*

Concatenative synthesis model is a kind of data-driven model. In this model, the voice of a sentence is generated by connecting small recorded voice units such as words, syllables, phones, diphones, triphones and so on. Then the synthesized utterance is modified by prosodic model. Because this method just changes the super-segmental features and the original segmental features remain unchanged maximally, its sound quality is very high.

It should be noted that the speech quality of this approach is affected by the unit length largely. With longer units, the less concatenation points are needed, and the naturalness increases, but the size of the speech corpus and the memory will become very numerous. With shorter units, less costs of memory and speech corpus are needed, but the sound quality will become bad. Therefore, constructing a good speech corpus is very important to concatenative synthesis.

The disadvantages of concatenative synthesis are the discontinuities of the unit boundaries and the artificial, stiff feeling of the prosody. Some unexpected errors will also cause the unstable results of the system. To these problems, an improvement scheme is unit selection synthesis.

D. *Unit Selection Synthesis*

For concatenative synthesis, the units must be modified by signal processing methods such as PSOLA to produce the prosody. A large modification often makes the speech sound unnatural. Unit selection synthesis is also a kind of concatenative synthesis, but it solves the problem of unnaturalness by storing multiple instances of each unit with varying prosodic features. The unit that matches closest to the target prosody will be selected and concatenated so that the prosodic features will not be modified largely and the sound quality will be improved.

The technique of unit selection synthesis was first introduced by Sagisaka et al. [4]. To choose the best units, prosodic features such as intonation and duration have been added to the target specification in CHATR system [5]. In order to solve the mis-matching problem in unit selection, the researchers proposed a harmonic plus noise model (HNM), which considers the speech signal as a weighted sum of the harmonics and noise of various components. This model makes the synthesized speech more natural [6].

However, this method also has the same problems of selecting error units and huge speech corpus as the common concatenative synthesis. In addition, the voice quality of this approach almost could not be changed.

E. *Hidden Markov model(HMM) Synthesis*

As mentioned above, unit selection synthesis needs a huge speech corpus. The database size increases in an enormous way and the training time is usually very long.

An alternative scheme is to use statistical parametric synthesis techniques to infer acoustic parameters from speech data. The advantages of this approach are: less memory is needed to store the parameters than to store the speech data; more variations are allowable. For example, the voice qualities,

speaking styles, or emotions can easily be modified by transforming HMM parameters [7].

The most used model of statistical parametric synthesis techniques is hidden Markov model (HMM). HMM consists of two phases, the training phase and the synthesis phase.

The training phase is similar to that used in speech recognition systems. At the training phase, which features should be trained for the models are decided. The frequently used features are MFCC and its dynamic features, LogF0 and its dynamic features. The features are extracted per-frame and are put in a feature vector. Then the feature vectors will be modeled by context-dependent HMMs (phonetic, linguistic, grammatical and prosodic contexts are taken into account) [8] for each phone.

At the synthesis phase, the feature vectors for a given phone sequence would be generated by content-dependent HMMs & duration models and a synthesis filter, such as STRAIGHT [9], would be used to transform the feature vectors into acoustic signals.

Fig. 6 shows a typical HMM-based synthesis system.

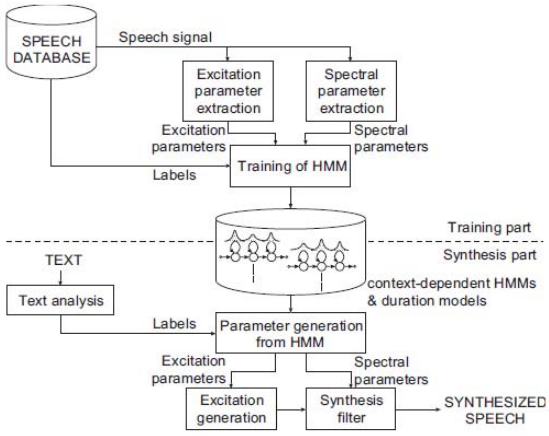

Fig. 6. A typical HMM-based speech synthesis system.

Based on the HMM model, the flexibility of this approach is higher, the speech corpus is smaller, and the training time is shorter than unit selection synthesis. However, HMM-based synthesis still has some disadvantages. The major limitation is the quality of the synthesized speech. Zen et al. [10] highlighted three major factors that degrade the quality of the synthesized speech: vocoding, accuracy of acoustic models, and over-smoothing.

F. Deep Neural Network Model

In HMM-based speech synthesis system, a number of contextual factors (around 50) should be taken into account in acoustic modeling. In order to cover all contextual factors with limited data, the top-down decision tree based context clustering is widely used, but it still has some limitations. First, it is inefficient to express complex context dependencies such as XOR, parity or multiplex problems. Second, the approach of dividing the input space and using separate parameters for each region would result in fragmenting the training data. These limitations would degrade the accuracy of acoustic models and quality of synthesized speech [11].

To address these limitations, an alternative scheme is to use DNN (Deep Neural Networks). DNN is a kind of multilayer neural network [12]. It will do non-supervised learning before supervised learning, and uses the weight values got from the former step as the initial values of the latter steps. The essence of depth learning is to learn more useful features by building a machine learning model with many hidden layers and big data training, eliminating the problems of manually selecting features. It can ultimately improve classification or prediction efforts.

In the speech synthesis, DNN could be used to models the features of text and speech sounds, and output the vectors of phonetic signals. Compared with HMM method, DNN has advantage at the classification of voice/voiceless and quasi-cycle prediction, but it costs more computation and the prediction of pitch is not good as HMM-based approach [11] [13]. According to these advantages and disadvantages, some researches to DNN have been carried on to improve its synthesis results [14, 15].

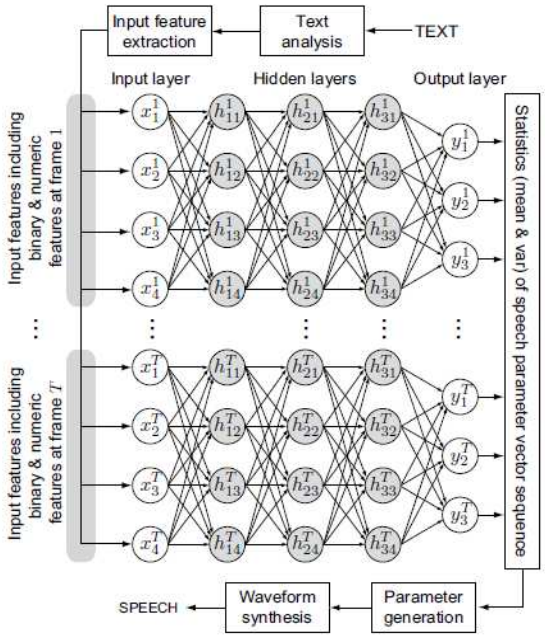

Fig. 7. A speech synthesis framework based on a DNN [11].

IV. CONCLUSION AND DISCUSSION

Finally, we would compare the speech synthesis approaches briefly. Rule-driven synthesis methods are used less today because the synthesized speech sounds are not good enough. But their advantages are the low processing costs and high flexibility. Data-driven synthesis methods nowadays are the dominant methods because they can provide higher quality of synthesized speech [7]. Unit selection synthesis could

generate the best speech but it also cost the largest memory and data. Compared with concatenative synthesis and unit selection synthesis, HMM, DNN and other statistical synthesis techniques could produce similar natural speech but only need less speech data and memory. They achieve the best balance at the present time.

From the viewpoint of technology, the integration of a variety of techniques will provide more improvement rooms for TTS techniques. Merging of rule-driven and data-driven approaches will be a possible direction of speech synthesis.

From the viewpoint of application, the demands to synthesized speech are more natural, personalized and emotional. Speech synthesis technology based on cloud source and embedded TTS systems would provide higher quality and more convenient for users, and they would become the hot spots of applications in the future. With the development of speech synthesis, the Text-to-Speech synthesis will play a greater role in social life.

REFERENCES

[1] Kroger B., "Minimal Rules for Articulatory Speech Synthesis.", In Proceedings of EUSIPCO92, pp.331-334, 1992.

[2] Klatt D. H., "Review of text-to-speech conversion for English." Journal of the Acoustical Society of America, vol. 82(3), 1987.

[3] Klatt D. H., "Software for a cascade/parallel formant synthesizer." Journal of the Acoustical Society of America, vol. 67, 1980.

[4] Sagisaka Y. et al., "ATR-TALK speech synthesis system." in Proceedings of International Conference on Spoken Language Processing, pp. 483–486, 1992.

[5] Black A. W. et al., "CHATR: A Generic Speech Synthesis System." In Proceedings of the International Conference on Computational Linguistics, pp. 983–986, 1994.

[6] Beutnagel M. et al., "The AT&T Next-Gen TTS System." in Proceedings of the 137th meeting of the Acoustical Society of America, 1999.

[7] M. Z. Rashadll. Et al., "An Overview of Text-To-Speech Synthesis Techniques", Latest trends on communications and information technology, 2010, pp 84-89.

[8] H. Kawahara, I. M asuda-Katsuse and A. deCheveigne, Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction:possible role of a repetitive structure in sounds, Speech Communication. 1999, vo .l 27, pp. 187 – 207.

[9] Heiga Zen, et al. The HMM-basedSpeech Synthesis System (HTS) Version 2.0, in Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.

[10] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," Speech Commun., vol. 51, no. 11, pp. 1039–1064, 2009.

[11] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proceedings of ICASSP, 2013, pp. 7962–7966.

[12] Zhou Zhi-hua, Chen Shi-fu. "Neural network ensemble". Chinese Journal of Computers, 2002, 25(1) : 1-8.

[13] Chen Zhen-huai, Yu Kai. "An investigation of implementation and performance analysis of DNN based speech synthesis system". in Proceedings of International Conference on Sociology and Psychology, pp. 577-582, 2014.

[14] Sankar Mukherjee, Shyamal Kumar Das Mandal. "F0 modeling in HMM-based speech synthesis system using deep belief network". Co-ordination and Standardization of Speech Database and Assessment Techniques, 2014, pp. 1-5.

[15] Yao Qian, Yuchen Fan, Frank K. Song. "On the training aspects of deep neural network(DNN) for parametric TTS synthesis". in Proceedings of International Conference on Acoustic, Speech and Signal Processing, pp. 3829-3833, 2014.