ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

UDK 004.934

Aitim A.K.*, Satybaldiyeva R.Zh.

*International Information Technology University, Almaty, Kazakhstan

ANALYSIS OF METHODS AND MODELS FOR AUTOMATIC PROCESSING SYSTEMS OF SPEECH SYNTHESIS

Abstract. *The article considers the current state of models and methods of speech synthesis. Their advantages and disadvantages are presented, as well as metrics for the quality of speech synthesis. The method of speech synthesis by rules is based on a programmed knowledge of acoustic and linguistic limitations and does not directly use elements of human speech.*

Key words: *Speech synthesis, speech recognition, model of speech synthesis, quality metrics, TTS.*

Introduction

Artificial creation of human speech has long been of interest to scientists and practical researchers. The tasks facing speech synthesizers have changed significantly over time. Speech synthesis has evolved from simply producing speech-like sounds and voicing a limited set of phrases to voicing any text with the desired intonation and emotional coloring, and in a voice, that accurately imitates the speech of a person. If we consider the synthesizer not just as a tool for pronouncing certain texts, but as part of the speech interface system for communication between a person and a computer, then it currently faces even more complex tasks. For example, the so-called ‘reactive’ synthesis must track the effect it has on the listener and change its characteristics accordingly. A pertinent problem today is also multi-language adaptive speech synthesis, that is, speech synthesis in different languages by the voice of a single speaker, who may not even know the target language.

Speech is the main means of communication between people. Speech synthesis, the automatic generation of speech signals, has been developed for several decades [1]. Subsequent advances in speech synthesis have created synthesizers with very high intelligibility, but sound quality and naturalness are still a serious problem. However, the quality of existing products has reached an adequate level for several applications, such as multimedia and telecommunications. With some audio-visual information or facial animation, speech intelligibility can be significantly improved [2].

The text-to-speech synthesis (TTS) procedure consists of two main stages. The first is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second is speech signal generation, when an acoustic output is created based on this phonetic and prosodic information. These two phases are usually referred to as high-level and low-level synthesis. A simplified version of the procedure is shown in Fig. 1. Input text can be, for example, data from a word processor, standard ASCII from an email, a mobile text message, or scanned text from a newspaper. The character string is then pre-processed and analyzed into a phonetic representation, which is usually a string of phonemes with some additional information for correct intonation, duration, and stress. The speech sound is finally generated using a low-level synthesizer based on information from a high-level synthesizer.

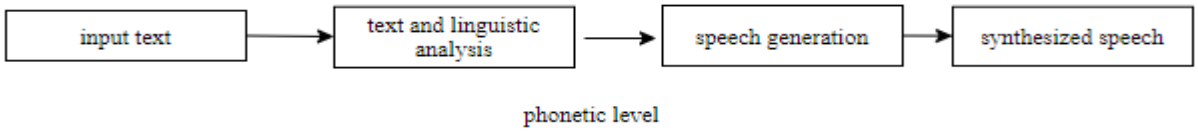

Figure 1 - Simple procedure to convert text to speech

The easiest way to create synthesized speech is to play long pre-recorded samples of natural speech, such as individual words or sentences. This method of combining provides high quality and naturalness but has a limited vocabulary and usually only one voice. The method is very suitable for some information and information systems. However, there is no database of all the words and common names in the world. It may even be inappropriate to call this speech synthesis because it contains only recordings. Thus, for unlimited speech synthesis (text-to-speech conversion), we must use shorter fragments of the speech signal, such as syllables, phonemes, diphones, or even shorter segments.

Another widely used method for creating synthesized speech is formant synthesis, which is based on the speech filter model of the production source. This method is sometimes called terminal analogy because it simulates only the sound source and formant frequencies, and not any physical characteristics of the vocal tract [3]. The excitation signal can be voiced with a basic frequency (F0) or non-vocalized noise. The mixed arousal of these two can also be used for voiced consonants and some aspiration sounds. The arousal is then obtained and filtered by a voice path filter that consists of resonators like natural speech formats.

Theoretically, the most accurate method for creating artificial speech is direct modeling of the human speech production system [4]. This method, called articulatory synthesis, typically involves models of the human articulators and vocal cords. Articulators are usually modeled using a set of functions for small tube sections. The vocal cord model is used to generate the corresponding excitation signal, which can be, for example, a two-mass model with two vertically moving masses [5]. Articulation synthesis promises high-quality synthesized speech.

All synthesis methods have their own advantages and problems, and it is quite difficult to say which method is the best. With concatenative and formant synthesis, very promising results have been achieved recently, but articulation synthesis may also emerge as a potential method in the future. Various synthesis methods, algorithms, and methods will be discussed in more detail later.

Implementation of speech synthesis

Speech synthesis is the creation of sound based on text. Speech synthesis may be required in all cases when the recipient of information is a person. The quality of a speech synthesizer is primarily judged by its similarity to the human voice, as well as its ability to be understood.

A wide variety of speech synthesis methods are currently available. There are two main factors that determine the choice of synthesis technology in an implementation:

1. Task. Depending on the requirements for the quality of the final product, the capabilities of synthesized speech vary. The simplest synthesized speech can be created by combining parts of recorded speech that will be stored in a database. Of course, if you need to synthesize a complex text, this method cannot be used, since at the junction of the composed sound fragments, there may be intonation distortions and breaks that are noticeable to the ear. In addition, you will need a very large database to store all the necessary audio fragments.

2. The structure of the language. The main phonological laws, stress rules, morphological and syntactic structures are used in constructing the output speech wave.

3. Technological capabilities. First, this is the amount of memory available for the information system. Depending on the amount of stored synthesized vocabulary, both its complexity and the quality of the resulting signal change. The computing power of the device plays an equally im-

portant role in choosing the method. Choosing a complex speech synthesis method, coupled with low hardware performance, will result in a huge amount of computing time.

Speech synthesizers are generally divided into two types: with a limited and unlimited dictionary. In devices with a limited dictionary, speech is stored in the form of words and sentences that are output in a certain sequence when synthesizing a speech message.

The main methods with a limited dictionary are the compilation synthesis model and parametric representation. [6]

2. Quality metrics

Before talking about which models of speech synthesis are better, we need to determine the quality metrics that will be used for comparing algorithms.

Since the same text can be read in an infinite number of ways, there is no a priori correct way to pronounce a phrase. Therefore, metrics of speech synthesis quality are often subjective and depend on the listener's perception.

The standard metric is the MOS (mean opinion score), an average rating of natural speech given by assessors for synthesized audio on a scale from 1 to 5. One means a completely improbable sound, and five means speech that is indistinguishable from human speech. Real people's records usually get values of about 4,5, and a value greater than 4 is considered high enough. Speech synthesis works as follows. The first step to building any speech synthesis system is to collect data for training. Usually these are high-quality audio recordings where the announcer reads specially selected phrases. The approximate size of the dataset required for training unit selection models is 10-20 hours of pure speech [7], while for neural network parametric methods, the upper estimate is approximately 25 hours [10, 11].

Today, speech synthesis problems are solved mainly using two approaches:

- Unit selection [8], or the compilation approach. It is based on merging fragments of recorded audio. Since the late 90's, it has long been considered the de facto standard for developing speech synthesis engines. For example, a voice using the unit selection method can be found in Siri [7].

- Parametric speech synthesis [9], the essence of which is to construct a probabilistic model that predicts the acoustic properties of an audio signal for a given text.

The speech of unit selection models (Fig. 2) is of high quality, low variability, and requires a large amount of data for training. At the same time, much less data is needed to train parametric models, they generate more diverse intonations. The compilation synthesis model assumes speech synthesis by concatenating recorded samples of individual sounds voiced by the speaker (Fig. 2).

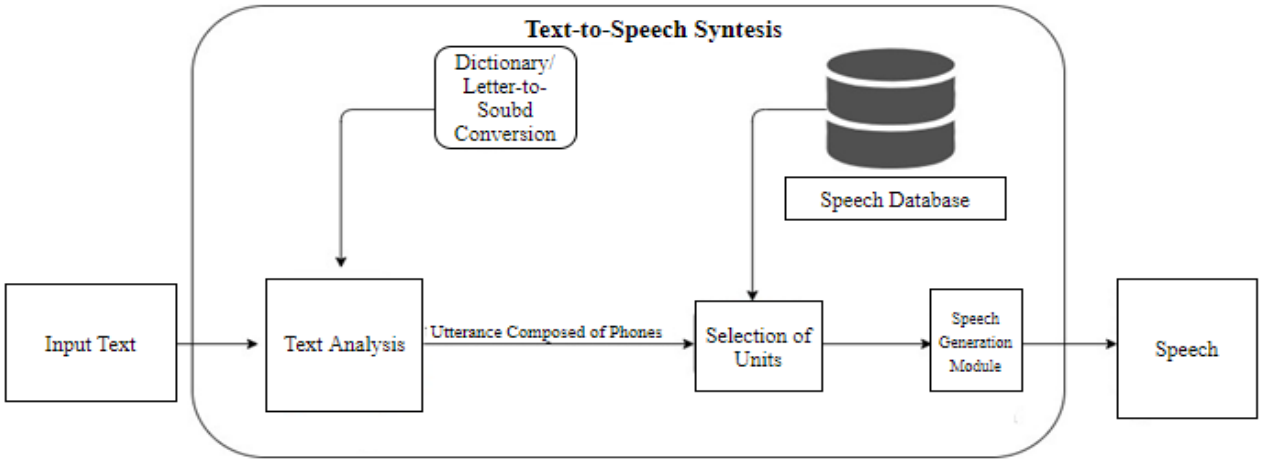

Figure 2 - Model of compiled synthesis or unit selection

The compilation synthesis model assumes speech synthesis by concatenating recorded samples of individual sounds uttered by the speaker.

When using this model, a database of audio fragments is compiled, from which speech will be synthesized in the future. The size of the synthesis elements is usually no less than a word.

Usually, the recorded speech of the speaker cannot cover all possible cases in which synthesis will be used. Therefore, the essence of the method is to split the entire audio database into small fragments called units, which are then glued together using minimal post-processing. The units are usually minimal acoustic units of the language, such as semitones or diphones [7].

Parametric speech synthesis

To solve two main problems of compilation synthesis, a parametric type of signal has been developed that is abstracted from the speech wave, which represents certain parameters. This approach reduces the amount of memory required for the dictionary and provides more flexibility compared to the compiled model.

Parameters display the most characteristic information or time or frequency zones. One way to configure it is to display the speech wave with the addition of individual harmonics at a given frequency.

Another variant of parametric vision of the speech path is artificial speech, creating the necessary set of formant resonances. This system works with the settings of the main tone and formants.

Formants are an economical way to store speech information, more than by reducing the required amount of memory compared to the compiled method.

The second advantage of this approach is its inherent flexibility. Semantic information consists of formants and the melody (intonation, speech dynamics, etc.) at the stage of the main tone and during the temporary division of speech, which allows you to divide what is pronounced and how the formant representation is pronounced. [9]

Thus, the formant approach requires less memory than the compiler, but it requires large calculations to reproduce the initial speech signal. Appropriate digital techniques and knowledge of speech formation models are required, but the linguistic structure of the language is not used.

The parametric approach is based on the idea of creating a probable model that estimates the propagation of acoustic features of a given text.

The full synthesis according to the rules

When synthesizing speech according to rules, the compilation and parametric encoding methods are also used, but at the syllable level.

In search of a compromise between the flexibility of full speech synthesis by rules and its cost-effectiveness, speech synthesis by rules using pre-memorized segments of natural language was developed.

This method is a variation of the conventional synthesis by the rules. Depending on the size of the initial synthesis elements, the following types of synthesis are distinguished: microsegment (microwave); allophonic; diphonic; semi-syllabic; syllabic; synthesis from units of any size.

Usually, these elements are used as semi-syllables-segments containing half of a consonant and half of a vowel adjacent to it.

The quality of this synthesis does not match the quality of natural speech, because distortions often occur at the borders of the cross-linking of diphones. Compiling speech from pre-recorded word forms also does not solve the problem of high-quality synthesis of arbitrary messages since words change depending on the type of phrase and the place of the word in the phrase. This position does not change even when using large amounts of memory to store word forms. However, this method of speech generation will give a higher quality sound output, compared to the simple method of synthesis by rules.

Conclusion

Today, speech synthesis is widely used in various areas of infrastructure. There are a small number of libraries for high-level programming languages that allow you to use the technologies

described above without resorting to a large amount of hardware power since the speech synthesis service itself is located on a cloud server. With the development of this direction, it will be almost impossible to distinguish artificial speech from a human speech in the future.

The paper considered the current state of models and methods of speech synthesis. Their advantages and disadvantages are presented, as well as metrics for the quality of speech synthesis. The method of speech synthesis by rules is based on a programmed knowledge of acoustic and linguistic limitations and does not directly use elements of human speech.

To remember this information requires little memory, but to extract parameters from it, you need the knowledge of an expert. Text analysis is a linguistic task and includes the definition of basic phonetic, syllabic, morphemic and syntactic forms, plus the extraction of semantic information. Text-to-speech conversion systems are the most complex speech synthesis systems that include knowledge about the structure of the human speech apparatus and the linguistic structure of the language.

Thus, this method gives complete freedom to model parameters and allows you to reproduce almost any text; it significantly saves memory, without requiring the storage of a large amount of information. However, synthesized speech sounds worse than natural speech (and, as a rule, worse than synthesized speech by other methods described above); such a system is difficult to develop.

REFERENCES

1. Möbius B., Schroeter J., Santen J., Sproat R., Olive J. Recent Advances in Multilingual Text-to-Speech Synthesis. Fortschritte der Akustik, DAGA-96, 2012
2. Beskow J. Talking Heads - Communication, Articulation, and animation. Proceedings of Fonetik-96: 53-56, 2012
3. Flanagan J. Speech Analysis, Synthesis, and Perception. Springer-Verlag, Berlin-Heidelberg-New York, 2011
4. O'Saughnessy D. Speech Communication - Human and Machine, Addison-Wesley, 2011
5. Veldhuis R., Bogaert I., Lous N. Two-Mass Models for Speech Synthesis. Proceedings of Eurospeech, 2009
6. Sorokin, V.N. Synthesis of speech. —M .: Nauka, 1992. - 392 p.
7. T. Capes, P. Coles, A. Conkie, L. Golipour, A. Hadjitarkhani, Q. Hu, N. Huddleston, M. Hunt, J. Li, M. Neeracher, K. Prahallad, T. Raitio, R. Rasipuram, G. Townsend, B. Williamson, D. Winarsky, Z. Wu, H. Zhang. Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System, Interspeech, 2017.
8. A.J. Hunt, A.W. Black. Unit selection in a concatenative speech synthesis system using a large speech database, ICASSP, 1996.
9. H. Zen, K. Tokuda, A. W. Black. Statistical parametric speech synthesis, Speech Communication, Vol. 51, no. 11, pp. 1039-1064, 2009.
10. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis.
11. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.
12. Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks.
13. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio.
14. Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg,

Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, Demis Hassabis. Parallel WaveNet: Fast High-Fidelity Speech Synthesis.

15. Wei Ping Kainan Peng Jitong Chen. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech.

Әйтім Ә.Қ.*, Сатыбалдиева Р.Ж.

Сөйлеу синтезінің автоматты өңдеу жүйелерінің әдістері мен үлгілерін талдау

Аңдатпа: Мақалада сөйлеу синтезінің үлгілері мен әдістерінің қазіргі жағдайы талқыланады. Олардың артықшылықтары мен кемшіліктері, сонымен қатар сөйлеу синтезі сапасының өлшемдері қарастырылған. Ережеге негізделген сөйлеу синтезінің әдісі акустикалық және тілдік шектеулер туралы бағдарламаланған ақпаратқа негізделген және адамның сөйлеу элементтерін тікелей пайдаланбайды.

Түйінді сөздер: Сөйлеу синтезі, сөйлеуді тану, сөйлеу синтезінің үлгісі, сапа көрсеткіштері, TTS.

Әйтім Ә.Қ.*, Сатыбалдиева Р.Ж.

Анализ методов и моделей систем автоматической обработки синтеза речи

Абстракт. В статье рассматривается современное состояние моделей и методов синтеза речи. Представлены их достоинства и недостатки, а также метрики качества синтеза речи. Метод синтеза речи по правилам базируется на запрограммированном знании акустических и лингвистических ограничений и не использует непосредственно элементы человеческой речи.

Ключевые слова: синтез речи, распознавание речи, модель синтеза речи, метрики качества, TTS.

Сведения об авторах:

Әйтім Әйгерім Қайратқызы, магистр технических наук, сениор-лектор кафедры «Информационных систем», Международный университет информационных технологий.

Сатыбалдиева Рысхан Жакановна, кандидат технических наук, ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

About authors:

Aigerim K. Aitim, master of technical sciences, senior-lecturer of the "Information Systems" department, International Information Technology University.

Ryskhan Zh. Satybaldiyeva, candidate of technical sciences, associate professor of the "Information Systems" department, International Information Technology University.