ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

A Review on Speech Synthesis an Artificial Voice Production

Smita S. Hande

Assistant professor, Dept. of ECE, Fr. C R I T, Sector 9A Vashi, Navi Mumbai, Maharashtra State, India

ABSTRACT: Speech is used to convey emotions, feelings and information. It has began from day one when human beings start to communicate. Speech synthesis, also called text-to-speech, is the generation of synthetic speech. An application or other process sends text to a speech synthesizer, which creates a spoken version that can be output through the audio hardware or saved to a file. This paper deals with some methods of speech synthesis. Speech synthesis has very wide range of applications, which are reviewed in later part of the paper.

KEYWORDS: Speech synthesis, Articulatory synthesizer, Formant synthesizer, Concatenative synthesizer.

I.INTRODUCTION

Speech synthesis is the artificial production of human speech [1]. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. The Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. The most important qualities of a speech synthesis system are naturalness and Intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech synthesizer is both natural and intelligible, hence speech synthesis systems usually try to maximize both characteristics.

Speech synthesized methods are usually classified into three groups:

- Articulatory synthesis, which attempts to model the human speech production system directly.
- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model [2-4].
- Concatenative synthesis, which uses different length prerecorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used methods. The articulatory method is complicated for high quality implementations.

Articulatory synthesis involves vocal cords and models of the human articulators with a set of area functions between glottis and mouth. Articulatory synthesis allow accurate modeling of transients due to abrupt area changes, whereas formant synthesis models only spectral behavior (O'Saughnessy 1987).

Formant synthesis is based on the source-filter-model of speech and most widely used synthesis method. Two basic structures used are, parallel and cascade. Combination of these can be used for better performance. Formant synthesis also provides infinite number of sounds. Three formants are generally required to produce intelligible speech and up to five formants to produce high quality speech.

II. SPEECH SYNTHESIS

It considers a network with N mobile unlicensed nodes that move in an environment according to some stochastic mobility models. It also assumes that entire spectrum is divided into number of M non-overlapping orthogonal channels having different bandwidth. The access to each licensed channel is regulated by fixed duration time slots. Slot timing is assumed to be broadcast by the primary system. Before transmitting its message, each transmitter node, which is a node with the message, first selects a path node and a frequency channel to copy the message. After the path and channel selection, the transmitter node negotiates and handshakes with its path node and declares the selected channel frequency to the path. The communication needed for this coordination is assumed to be accomplished by a fixed length frequency hopping sequence (FHS) that is composed of K distinct licensed channels. In each time slot, each node consecutively hops on FHS within a given order to transmit and receive a coordination packet. The aim of coordination packet that is generated by a node with message is to inform its path about the frequency channel decided for the message copying.

This section covers the some important points of speech synthesis.

A. *ARTICULATORY* SPEECH SYNTHESIZER

Articulatory synthesis is the production of speech sounds using a model of the vocal tract, which directly or indirectly simulates the movements of the speech articulators. It provides a means for gaining an understanding of speech production and for studying phonetics. In such a model coarticulation effects arise naturally, and in principle it should be possible to deal correctly with glottal source properties, interaction between the vocal tract and the vocal folds, the contribution of the subglottal system, and the effects of the nasal tract and sinus cavities. Articulatory synthesis usually consists of two separate components. In the articulatory model, the vocal tract is divided into many small sections and the corresponding cross-sectional areas are used as parameters to represent the vocal tract characteristics. In the acoustic model, each cross-sectional area is approximated by an electrical analog transmission line. To simulate the movement of the vocal tract, the area functions must change with time. Each sound is designated in terms of a target configuration and the movement of the vocal tract is specified by a separate fast or slow motion of the articulators. A properly constructed articulatory synthesizer is capable of reproducing all the naturally relevant effects for the generation of fricatives and plosives, modeling coarticulation transitions as well as source-tract interaction in a manner that resembles the physical process that occurs in real speech production.

Articulatory models can be classified into two major types: parametric area model and midsagittal distance model shown in Fig. 1. The parametric area model describes the area function as a function of distance along the tract, subject to some constraints. The area of the vocal tract is usually represented by a continuous function such as a hyperbola, a parabola, or a sinusoid. The midsagittal distance model describes the speech organ movements in a midsagittal plane and specifies the position of articulatory parameters to represent the vocal tract shape. Coker and Fujimura (1966) introduced an articulatory model with parameters assigned to the tongue body, tongue tip, and velum. Later this model was modified to control the movements of the articulators by rules.

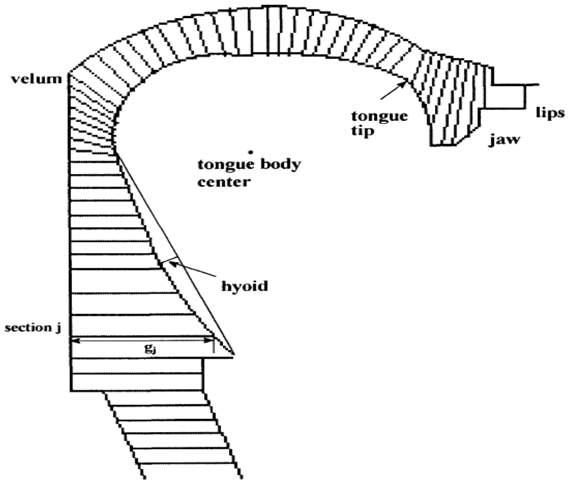

Fig. 1 Articulatory model's parameters and midsagittal grids.

B. Formant Synthesis:

Rule-based formant synthesis is based on a set of rules used to determine the parameters necessary to synthesize a desired utterance using a formant synthesizer (Allen et al. 1987). The input parameters may be for example the following, where the open quotient means the ratio of the open-glottis time to the total period duration (Holmes et al. 1990):

- Voicing fundamental frequency (F0)
- Voiced excitation open quotient (OQ)
- Degree of voicing in excitation (VO)
- Formant frequencies and amplitudes (F1...F3 and A1...A3)
- Frequency of an additional low-frequency resonator (FN)
- Intensity of low- and high-frequency region (ALF, AHF)

A cascade formant synthesizer (Fig. 2) consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the following one. The cascade structure needs only formant frequencies as control information. The main advantage of the cascade structure is that the relative formant amplitudes for vowels do not need individual controls (Allen et al. 1987).

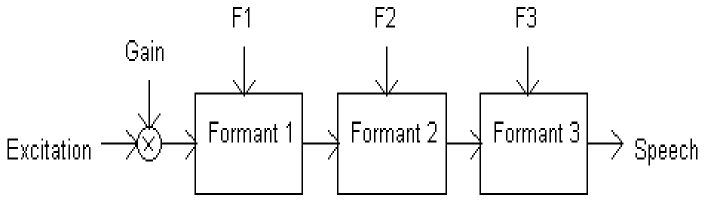

Fig. 2. Basic structure of cascade formant synthesizer.

The cascade structure shown in Fig. 2 has been found better for non-nasal voiced sounds and because it needs less control information than parallel structure, it is then simpler to implement. However, with cascade model the generation of fricatives and plosive bursts is a problem.

A parallel formant synthesizer Fig. 3 consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. The excitation signal is applied to all formants simultaneously and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or antiresonances in the frequency response (O'Saughnessy 1987). The parallel structure enables controlling of bandwidth and gain for each formant individually and thus needs also more control information.

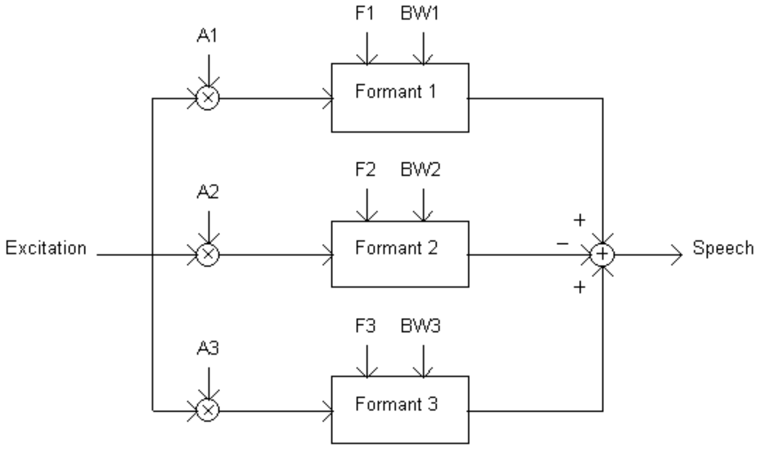

Fig. 3 Basic structure of a parallel formant synthesizer.

The parallel structure has been found to be better for nasals, fricatives, and stop-consonants, but some vowels can not be modeled with parallel formant synthesizer as well as with the cascade one.

There has been widespread controversy over the quality and suitably characteristics of these two structures. It is easy to see that good results with only one basic method is difficult to achieve so some efforts have been made to improve and combine these basic models. In 1980 Dennis Klatt (Klatt 1980) proposed a more complex formant synthesizer which incorporated both the cascade and parallel synthesizers with additional resonances and anti-resonances for nasalized sounds, sixth formant for high frequency noise, a bypass path to give a flat transfer function, and a radiation characteristics. The system used quite complex excitation model which was controlled by 39 parameters updated every 5 ms. The quality of Klatt Formant Synthesizer was very promising and the model has been incorporated into several present TTS systems.

C. *Concatenative Synthesis*

Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods.

One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units. With longer units high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems units used are usually words, syllables, demisyllables, phonemes, diphones, and sometimes even triphones.

Word is perhaps the most natural unit for written text and some messaging systems with very limited vocabulary. Concatenation of words is relative easy to perform and coarticulation effects within a word are captured in the stored units. However, there is a great difference with words spoken in isolation and in continuos sentence which makes the continuous speech to sound very unnatural (Allen et al. 1987). Because there are hundreds of thousands of different words and proper names in each language, word is not a suitable unit for any kind of unrestricted TTS system.

The number of different syllables in each language is considerably smaller than the number of words, but the size of unit database is usually still too large for TTS systems. For example, there are about 10,000 syllables in English. Unlike with words, the coarticulation effect is not included in stored units, so using syllables as a basic unit is not very reasonable. There is also no way to control prosodic contours over the sentence. At the moment, no word or syllable based full TTS system exists. The current synthesis systems are mostly based on using phonemes, diphones, demisyllables or some kind of combinations of these.

Demisyllables represents the initial and final parts of syllables. One advantage of demisyllables is that only about 1,000 of them is needed to construct the 10,000 syllables of English (Donovan 1996). Using demisyllables, instead of for example phonemes and diphones, requires considerably less concatenation points. Demisyllables also take account of most transitions and then also a large number of coarticulation effects and also covers a large number of allophonic variations due to separation of initial and final consonant clusters. However, the memory requirements are still quite high, but tolerable. Compared to phonemes and diphones, the exact number of demisyllables in a language can not be defined. With purely demisyllable based system, all possible words can not be synthesized properly. This problem is faced at least with some proper names (Hess 1992). Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech. The inventory of basic units is usually between 40 and 50, which is clearly the smallest compared to other units (Allen et al. 1987). Using phonemes gives maximum flexibility with the rule-based systems. However, some phones that do not have a steady-state target position, such as plosives, are difficult to synthesize. The articulation must also be formulated as rules. Phonemes are sometimes used as an input for speech synthesizer to drive for example diphone-based synthesizer.

Diphones (or dyads) are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transitions between adjacent phones. That means that the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation points. Another advantage with diphones is that the coarticulation effect needs no more to be formulated as rules. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all combinations of phonemes are needed. For example, in Finnish the combinations, such as /hs/, /sj/, /mt/, /nk/, and /h p/ within a word are not possible. The number of units is usually from 1500 to 2000, which increases the memory requirements and makes the data collection more difficult compared to phonemes. However, the number of data is still tolerable and with other advantages, diphone is a very suitable unit for sample-based text-to-speech synthesis. The number of diphones may be reduced by inverting symmetric transitions, like for example /as/ from /sa/.

Longer segmental units, such as triphones or tetraphones, are quite rarely used. Triphones are like diphones, but contains one phoneme between steady-state points (half phoneme - phoneme - half phoneme). In other words, a triphone is a phoneme with a specific left and right context. For English, more than 10,000 units are required (Huang et al. 1997).

Building the unit inventory consists of three main phases (Hon et al. 1998). First, the natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the units must be labeled or segmented from spoken speech data, and finally, the most appropriate units must be chosen. Gathering the samples from natural speech is usually very time-consuming. However, some is this work may be done automatically by choosing the input text for analysis phase properly. The implementation of rules to select correct samples for concatenation must also be done very carefully.

There are several problems in concatenative synthesis compared to other methods.

- Distortion from discontinuities in concatenation points, which can be reduced using diphones or some special methods for smoothing signal.
- Memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labeling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.
- Some of the problems may be solved with methods described below and the use of concatenative method is increasing due to better computer capabilities (Donovan 1996).

III. APPLICATIONS OF SPEECH SYNTHESIS

In this scheme, each node with message searches for possible path nodes to copy its message. Hence, possible path nodes of a node are considered. Using NSS, each node having message selects its path nodes to provide a sufficient level of end-to-end latency while examining its transmission effort. Here, it derives the CSS measure to permit CR-Networks nodes to decide which licensed channels should be used. The aim of CSS is to maximize spectrum utilization with minimum interference to primary system. Assume that there are M licensed channels with different bandwidth values and y denotes the bandwidth of channel c. Each CR-Networks node is also assumed to periodically sense a set of M licensed channels. Mi denotes the set including Ids of licensed channels that are periodically sensed by node i. suppose that channel c is periodically sensed by node i in each slot and channel c is idle during the time interval x called channel idle duration. Here, it use the product of channel bandwidth y and the channel idle duration x, tc = xy, as a metric to examine the channel idleness. Furthermore, failures in the sensing of primary users are assumed to cause the collisions among the transmissions of primary users and CR-Networks nodes.

Augmented communicators—individuals who cannot produce understandable speech and instead use synthetic speech generated by an Augmentative and Alternative Communication (AAC) device—have for years relied on a small number of commercially available synthetic “voices” for use in their AAC devices. Mostly, these devices have used rule-based formant synthesis systems to generate synthetic speech. Thus, many AAC devices have relied upon synthesis technology that is decades old and demonstrably less intelligible and less natural sounding. Some applications are listed in this section,

A. Applications for the Blind

The most important and useful application field in speech synthesis is the reading and communication aids for the blind. Before synthesized speech, specific audio books were used where the content of the book was read into audio tape. It is also easier to get information from computer with speech instead of using special bliss symbol keyboard, which is an interface for reading the Braille characters.

The first commercial TTS application was probably the Kurzweil reading machine for the blind introduced by Raymond Kurzweil in the late 1970's. A speech synthesizer will be very helpful and common device among visually impaired people in the future. Current systems are mostly software based, so with scanner and OCR system, it is easy to construct a reading machine for any computer environment with tolerable expenses. Regardless of how fast the development of reading and communication aids is, there is always some improvements to do.

The most crucial factor with reading machines is speech intelligibility which should be maintained with speaking rates ranging from less than half to at least three times normal rate (Portele et al. 1996). Naturalness is also an important feature and makes the synthetic speech more acceptable. Although the naturalness is one of the most important features, it may sometimes be desirable that the listener is able to identify that speech is coming from machine (Hess 1992), so the synthetic speech should sound natural but somehow "neutral".

When the output from a speech synthesizer is listened for the first time, it may sound intelligible and pleasant. However, during longer listening period, single clicks or other weak points in the system may arise very annoying. This is called an annoying effect and it is difficult to perceive with any short-term evaluation method, so for these kind of cases, the feedback from long-term users is sometimes very essential.

Speech synthesis is currently used to read www-pages or other forms of media with normal personal computer. Information services may also be implemented through a normal telephone interface with keypad-control similar to text-tv. With modern computers it is also possible to add new features into reading aids. It is possible to implement software to read standard check forms or find the information how the newspaper article is constructed. However, sometimes it may be impossible to find correct construction of the newspaper article if it is for example divided in several pages or has an anomalous structure.

A blind person can not also see the length of an input text when starting to listen it with a speech synthesizer, so an important feature is to give in advance some information of the text to be read. For example, the synthesizer may check the document and calculate the estimated duration of reading and speak it to the listener. Also the information of bold or underlined text may be given by for example with slight change of intonation or loudness.

B. Applications for the Deafened and Vocally Handicapped

People who are born-deaf can not learn to speak properly and people with hearing difficulties have usually speaking difficulties. Synthesized speech gives the deafened and vocally handicapped an opportunity to communicate with people who do not understand the sign language. With a talking head it is possible to improve the quality of the communication situation even more because the visual information is the most important with the deaf and dumb. A speech synthesis system may also be used with communication over the telephone line (Klatt 1987).

With keyboard it is usually much slower to communicate than with normal speech. One way to speed up this is to use the predictive input system that always displays the most frequent word for any typed word fragment, and the user can then hit a special key to accept the prediction. Even individual pre-composed phrases, such as greetings or salutes, may be used.

C. Educational Applications

Synthesized speech can be used also in many educational situations. A computer with speech synthesizer can teach 24 hours a day and 365 days a year. It can be programmed for special tasks like spelling and pronunciation teaching for different languages. It can also be used with interactive educational applications.

Especially with people who are impaired to read (dyslexics), speech synthesis may be very helpful because especially some children may feel themselves very embarrassing when they have to be helped by a teacher (Klatt 1987). It is also almost impossible to learn write and read without spoken help. With proper computer software, unsupervised training for these problems is easy and inexpensive to arrange.

A speech synthesizer connected with word processor is also a helpful aid to proof reading. Many users find it easier to detect grammatical and stylistic problems when listening than reading. Normal misspellings are also easier to detect.

D. Applications for Telecommunications and Multimedia

The newest applications in speech synthesis are in the area of multimedia. Synthesized speech has been used for decades in all kind of telephone enquiry systems, but the quality has been far from good for common customers. Today, the quality has reached the level that normal customers are adopting it for everyday use.

Electronic mail has become very usual in last few years. However, it is sometimes impossible to read those E-mail messages when being for example abroad. There may be no proper computer available or some security problems exists. With synthetic speech e-mail messages may be listened to via normal telephone line. Synthesized speech may also be used to speak out short text messages (sms) in mobile phones.

For totally interactive multimedia applications an automatic speech recognition system is also needed. The automatic recognition of fluent speech is still far away, but the quality of current systems is at least so good that it can be used to give some control commands, such as yes/no, on/off, or ok/cancel.

E. Other Applications and Future Directions

In principle, speech synthesis may be used in all kind of human-machine interactions. For example, in warning and alarm systems synthesized speech may be used to give more accurate information of the current situation. Using speech instead of warning lights or buzzers gives an opportunity to reach the warning signal for example from a different room. Speech synthesizer may also be used to receive some desktop messages from a computer, such as printer activity or received e-mail.

In the future, if speech recognition techniques reach adequate level, synthesized speech may also be used in language interpreters or several other communication systems, such as videophones, videoconferencing, or talking mobile phones. If it is possible to recognize speech, transcribe it into ASCII string, and then resynthesize it back to speech, a large amount of transmission capacity may be saved. With talking mobile phones it is possible to increase the usability considerably for example with visually impaired users or in situations where it is difficult or even dangerous to try to reach the visual information. It is obvious that it is less dangerous to listen than to read the output from mobile phone for example when driving a car.

During last few decades the communication aids have been developed from talking calculators to modern three-dimensional audiovisual applications.

VI. CONCLUSION

In this paper, the three main approaches to speech synthesis are described. While formant and articulatory syntheses offer perhaps more flexibility than concatenative synthesisers. The rise of concatenative synthesis began in the 70s, and has largely become practical as large-scale electronic storage has become cheap and robust.

Synthetic speech may be used in several applications. Communication aids have developed from low quality talking calculators to modern 3D applications, such as talking heads. The implementation method depends mostly on used application. In some cases, such as announcement or warning systems, unrestricted vocabulary is not necessary and the best result is usually achieved with some simple messaging system. With suitable implementation some funds may also be saved. On the other hand, some applications, such as reading machines for the blind or electronic-mail readers, require unlimited vocabulary and a TTS system is needed.

REFERENCES

[1] Rabiner, L. R., and Schafer, R. W., “Digital Processing of Speech Signals”, Prentice- Hall, Englewood Cliffs, NJ, 1978.

[2] Klatt, D. H., and Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers," The Journal of the Acoustical Society of America, vol. 87, pp. 820-857, 1990.

[3] Klatt, D. H., "Review of text-tospeech conversion for English," Journal of the Acoustical Society of America, vol. 82, pp. 737-793, 1987.

[4] Klatt, D. H., “Software for a Cascade/Parallel Formant Synthesiser”, The Journal of the Acoustical Society of America, 67(3), Mar. 1980, 971-995, 1980.

[5] Coker, C. H., "A model of articulatory dynamics and control," Proc. IEEE, 64(4),1976, 452-460.

[6] Mermelstein, P., "Articulatory model for the study of speech production," J. Acoust. Soc. Am., 53(4), 1973, 1070-1082.

[7] Sondhi, M. M. and Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer, " IEEE Trans. Acoust., Speech, and Signal Processing, 35(7), 1987, 955-967.

[8] Black, A., and Lenzo, K., “Limited domain synthesis,” in ICSLP2000, Beijing, China., 2000, vol. II, pp. 411–414.

[9] Kain, A., and Macon, M.,” Spectral voice conversion for text-to-speech synthesis”, In: Proc.ICASSP, Seattle,1998.