2007-01-01

A Brief Introduction to Speech Synthesis and Voice Modification

Alan O'Cinneide
Technological University Dublin, alan.ocinneide@tudublin.ie

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

Mikel Gainza
Technological University Dublin, Mikel.Gainza@tudublin.ie

Recommended Citation

O'Cinneide, A., Dorran, D. & Gainza, M. A Brief Introduction to Speech Synthesis and Voice Modification. *Sounds Electric 2007 Conference, Dundalk Institute of Technology.*

A Brief Introduction to Speech Synthesis and Voice Modification

Alan O Cinneide, David Dorran and Mikel Gainza
Audio Research Group, Dublin Institute of Technology
Kevin Street, Dublin 2, Ireland

ABSTRACT

For both engineers and linguists, the computer synthesis of natural speech is an objective that would provide many useful applications to human-computer interaction, including the realm of electro-acoustic music. The purpose of this paper is to introduce the area of speech synthesis by providing an overview of the three main methods of computer speech synthesis; namely concatenative, articulatory and formant syntheses. Some aspects of the current state of the technology are illuminated and the final section will explain the author's motivation and current research approach to the field of voice modification.

I. INTRODUCTION

The automatic conversion of written text into speech would prove a vitally useful technology with many commercial and humanitarian applications. Speech is one of the most natural ways for humans to communicate with one another; enabling a computer to convey information in this manner has marked advantages over written text.

Speech synthesis would be valuable as an assistive technology but it can also be applied to more general commercial applications. Amongst the typical uses of the technology are:

Talking aids for the vocally handicapped.
Any person who is unable to speak but can use a typewriter or similar interface has the potential to use a text-to-speech system to provide themselves with a voice. The synthesised voice can be tailored to the person with specific characteristics to retain individuality [1].

Reading aids for the blind.
The visually impaired can benefit tremendously from text-to-speech technology. Text-to-speech software would enable input text to be generated to spoken words, and allow access to information available online [1].

Training and educational aids.
With regard to many cognitive activities, it is known that speech has several advantages over written language. Virtual teachers, contributing to a distance learning course, for example, could teach on-line tutorials. This can be particularly advantageous in situations where the presence of a real teacher can be embarrassing for the student, as has been noted for sufferers of dyslexia [2].

Remote access to online information.
Any written information that is stored online, for example electronic mail, news items, directory enquiries, can all be accessed aurally by means of a speech synthesiser [1].

Firstly, this paper provides a general introduction to the problems involved with speech synthesis. The second section illustrates the three main methods by which it is accomplished by modern text-to-speech systems. Finally, the area of voice modification is introduced and the future work of the author briefly discussed.

II. TEXT-TO-SPEECH CONVERSION: PROCESS OUTLINE

The usual process of text-to-speech conversion is illustrated in Fig. 1.

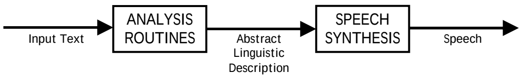

Fig. 1. The two step text-to-speech flowchart of operations.

The initial stage of the process consists of a set of analyses derived from computational linguistics. To obtain high-quality speech synthesis, lexicographical, syntactical and semantic analyses are required; the same knowledge is necessary for language translation [3]. The processes identify the words in the text and establish pronunciation and prosodic characteristics of the phrase, as stipulated by a set of pre-determined linguistic rules.

The result from this complex initial step is an abstract linguistic description of the utterance to be synthesised, consisting of phonemes, stress marks and syntactic structure indicators. This abstraction is then passed into the speech synthesiser where it is converted to control parameters which are then used to drive a speech synthesiser.

III. SPEECH SYNTHESIS

In latter part of the 18th century, a Hungarian nobleman, Wolfgang von Kempelen, developed a speaking machine based upon his observations of human speech production [3]. The machine consisted of a pressure chamber in place of the lungs, a vibrating metal reed representing the vocal folds, and a leather sack for the vocal tract, which could be manipulated to produce different vowel sounds. The inclusion of various different models of the tongue and lips made it possible to produce plosives. It was reported that the talking machine could speak whole phrases in French and Italian.[1]

Before the advent of electronic systems and computers, it was only through mechanical means like Kempelen's that speech-like sounds could be synthesised. Since the first third of the 20th century, the deeper understanding of the human speech production system coupled with technological advancements allowed for the development of electronically-based speech synthesis. There exist two broad categories of these approaches: rule-based and data-driven methods. Rule-based speech synthesis relies on some knowledge of the human speech system to help to assemble a model that can re-create those sounds. In contrast, data-driven synthesisers essentially ignore the speech production system and function by stringing together recorded speech segments.

III.1 RULE-BASED METHODS: THE SOURCE-FILTER THEORY OF SPEECH GENERATION

In order to develop a rule-based method of speech synthesis, a critical step is a description of the inner mechanisms of the human vocal system. This advance was taken with the proposal of the source-filter theory of speech production [4].

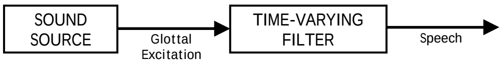

Fig. 2. The source-filter model of human speech production.

In this simplified approximation, the theory states that the human vocal system is described as a linearly connected two-part system: a source and a filter (Fig. 2). As air from the lungs passes through the vocal folds within the larynx, a pitched or un-pitched phonation occurs. This sound is seen as the source entering into the vocal tract. The vocal tract acts as a filter and alters the frequency content of the source sound, depending on the position of the jaw, tongue and articulators.

The source-filter theory is a simplification of the intricate relationship between the glottal source and the vocal tract. In actuality, they interact in a complex non-linear fashion that has yet to be satisfactorily described. A detailed theory will account for the effects of sub-glottal coupling, the subtleties of vocal fold motion and other speech production intricacies [4]. Despite its simplifications, the source-filter theory forms the basis for rule-based speech synthesis systems that can yield high quality results.

III.2 FORMANT SYNTHESIS

Spectrograms taken of speech signal reveal the dynamic nature of the frequency content of human speech. As speech is produced, the movements of the jaw, tongue and other articulators enhance certain resonances that change according to the vocal tract geometry. These resonant regions exhibited by the vocal tract are called formants[2]. Formant synthesis seeks to mimic human speech by artificially creating the movements of these frequency resonances.

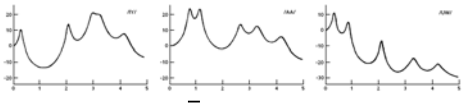

Fig. 3. Sample formant structures for the vowel /IY/, /AA/ and /UW/. Adapted from [5].

The parameters necessary for driving formant synthesis models can be derived from mathematical analyses of continuous speech, for example by Fourier analysis or by linear prediction methods [4].

Though the theory upon which it is based is a simplification of the actual human speech

[1] Von Kempelen's scientific reputation would have been assured by the unveiling of his invention, but his credibility was destroyed after his chess-playing automaton was revealed to be a hoax. The body of the automaton actually contained a midget who could control the movement of the pieces by the use of magnets.

[2] The word formant has its roots in the Latin verb *formāre*, meaning to shape or form.

production process, by making functional approximations to these phenomena, formant synthesisers can produce very high quality speech. Indeed, in the early 70s, Holmes showed that his formant synthesiser could be used to reproduce a nearly perfect duplicate of the male voice[3], but the determination of the complex control tables took months of manual adjustment [1].

A significant disadvantage afflicts formant synthesisers in that the dynamically changing formant frequencies and bandwidths bare no simple relationship with the articulatory specifications of the vocal tract [6]. Some modern formant synthesisers, for example Sensimetrics' HLsyn described in [7], attempt to negate this drawback by adopting a "quasi-articulatory" synthesis approach and translating human articulatory movements into the abstract formant synthesis control parameters through a set of mapping functions.

III.3 ARTICULATORY SYNTHESIS

While formant synthesis attempts to create the spectral features of a particular speech sound directly, articulatory synthesis does so by modelling the geometry of the vocal tract that would re-create a specific spectrum. By approaching speech production from this physical modelling standpoint, it has a direct relation to the human vocal system. Thus, it has been conjectured that this method will eventually lead to a complete vocal structure model, capable of reproducing all the sounds utter-able by the human speech reproduction system [6] [8].

Articulatory synthesis completes a full circle with the type of approach taken by von Kempelen hundreds of years ago. However, instead of physically manipulating a mechanical model of the human speech production system to produce utterances, speech is calculated as the output of a virtual model simulated on a computer. Typically, the vocal tract is divided into many small sections whose dimensions collectively determine the resonant characteristics of the vocal tract (Fig. 4). Mimicry of speech is achieved by dynamically changing the virtual shape and sizes of these segments according to the corresponding articulatory movements.

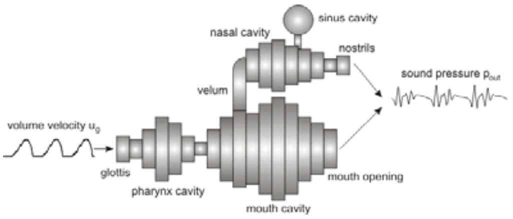

Fig. 4. A articulatory model of the human speech production system (Image taken from http://wwwicg.informatik.uni-rostock.de/~piet/speak_main.html).

Vocal tract dimensions have been obtained by measurements of x-rays and other special laboratory methods, but ideally these parameters would be derived from natural speech, in a manner similar to the parameter determination of formant synthesisers. This, however, is a non-trivial problem, afflicted by problems of articulatory ambiguity [8]. As such, a method to obtain articulatory parameters from a speech signal remains an on-going research direction [9].

III.4 DATA-BASED METHODS: CONCATENATIVE SYNTHESIS

It is possible to reproduce speech messages by playing back recordings of spoken words in the correct order to generate the desired message [10]. Such data-driven systems are called concatenative synthesisers, as they concatenate speech segments to produce utterances. This was the type of method employed by the UK telephone network's speaking clock, introduced in 1936. The various phrases and words were carefully recorded with regards to pitch, stress and other prosodic elements to ensure that reasonable fluency and naturalness are retained at synthesis time [3]. Similar schemes have been employed by transportation networks, meteorological services and other such systems requiring a limited set of words.

The concatenation of whole words can produce extremely natural sounding results but they can become restricted by their lack of flexibility. They are limited by the memory available for storage but also by the problems inherent in recording and editing the new words. However it can be possible to overcome these difficulties by concatenating speech segments which are less than a word in length.

The size of sub-word unit to use for purposes of concatenation needs to be chosen with care. Experiments using individual phones as the building blocks of speech demonstrated how normal speech sounds are heavily influenced by its neighbouring utterances. The co-articulation phenomenon results from the movement of the articulators approximating

[3] Holmes' attempts to model a female voice in a similar fashion wasn't as successful, indicating that the assumptions of the source filter model are less appropriate for the female speech production system.

target positions rather than exactly reaching them. Recordings played back with any regard for this co-articulation have been judged to by extremely difficult to listen to, as a result of the wrong intonation and rhythm [10].

As speech sounds generally consist of a steady-state region as well as transitional sections, researchers have been able to overcome the problem of co-articulation somewhat by concatenating phonemic transitions. However, there still remains a high degree of variation at the steady-state regions in these segments, according to the identity of the adjacent phones.

Modern concatenative systems utilise extremely large speech corpuses from which to draw their speech segments. Such systems are called unit-selection concatenative synthesisers, emphasising perhaps that the key to synthesis isn't the concatenating of speech segments but the selection of segments with the minimum of joins necessary, (Fig. 5). Specially designed cost functions determine which segments are chosen [11]. By reducing the amount of digital manipulation of the signal, a high degree of naturalness can be achieved.

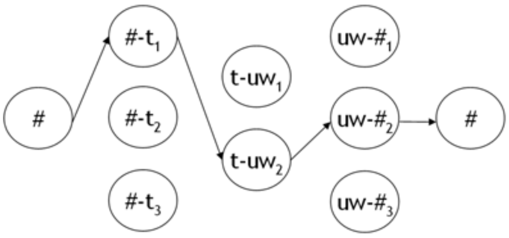

Fig. 5. A schematic representation of a unit selecting concatenative speech synthesiser.

More recently, in [8] it has been suggested that as concatenative systems essentially ignore any information about the speech production mechanism, improved systems can be devised if the method of concatenative synthesis enters the articulatory domain. By eliminating articulatory discontinuities, it is hoped the imperfect joins that persist due to co-articulation can be rectified and more natural results be achieved.

IV. VOICE MODIFICATION

Concatenative synthesis offers perhaps the most natural sounding computer speech. However, this may undesirably limit the flexibility of such as system because it cannot produce anything beyond the recorded material of the corpus. By implementing a voice modification module as the last stage of text-to-speech synthesis, a larger range of variation can be obtained from the corpus of a single speaker [12].

Other uses also exist. A voice modifier could be used to manipulate voices in such a way that the speaker's age or gender is disguised. Research has also been taken into the area of voice conversion for cross-language film dubbing [13].

In [14], it was noted that two acoustic parameters were perceptually important indicators of vocal texture: the degree of aspiration noise intruding at high frequencies in vowels, and the relative strength of the fundamental component of the glottal source wave. This justifies the assumption that in order to alter the vocal texture of an utterance, the voiced speech segments need only be manipulated. Thus, by changing the glottal source during these specific times and altering according to some specific guidelines, it is hoped that the vocal texture of any recorded utterance can be modified.

Gaining access to the glottal source involves the undoing of the effects of the vocal tract filter in an operation known as inverse-filtering. Typically, this mathematical procedure involves a time-series analysis technique called linear prediction. Using linear prediction methods, the resonances of the vocal tract filter can be removed from a speech signal, yielding the approximate glottal source. The results of the inverse-filtering operation depend heavily on the method of linear prediction and its parameters; numerous methods have been attempted in [15] and others.

Once the glottal waveform has been isolated, in order to alter its characteristics meaningfully, the glottal source model like the Liljencrants-Fant (LF) glottal model can be fit to the signal, on a period-by-period basis. The LF model is a well researched glottal model that has been shown to accommodate a large range of natural variation [16]. By changing the parameters accordingly, following, for example, the results of the experiments in [17], manipulating the glottal waveform to change the perceived vocal texture should be trivial.

V. 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, data-based methods are preferred because of their level of naturalness, seemingly unattainable by the rule-based versions. It is the author's opinion, concurring with those previously referred to in [6] and [8], that rule-based synthesis will eventually become the preferred method by which speech is synthesised, though this will only be realised when a more comprehensive speech production theory is proposed.

Also outlined is the author's research approach to the area of voice modification. Goals have been set to implement the voice quality modification procedure described. Inverse filtering and glottal model fitting techniques are currently being explored with a view to improving them.

REFERENCES

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

[2] R. C. Atkinson, "Teaching Children to Read Using a Computer," *American Psychologist*, vol. 29, pp. 169-178, 1974.

[3] M. R. Schroeder, *Computer Speech: Recognition, Compression, Synthesis*. Berlin: Springer-Verlag, 1999.

[4] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Engelwood Cliffs, New Jersey: Prentice-Hall Inc., 1978.

[5] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *The Journal of the Acoustical Society of America*, vol. 67, pp. 971-995, 1980.

[6] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. New York: John Wiley & Sons, Inc., 2000.

[7] H. M. Hanson and K. N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," *The Journal of the Acoustical Society of America*, vol. 112, pp. 1158, 2002.

[8] M. Mohan Sondhi, "Articulatory modeling: a possible role in concatenative text-to-speech synthesis," presented at Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on, 2002.

[9] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 133-150, 1994.

[10] J. N. Holmes, *Speech Synthesis and Recognition*. Berkshire: Van Nostrand Reinhold (UK) Co. Ltd., 1988.

[11] A. J. Hunt and A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, Georgia, 1996.

[12] D. Sündermann, "Voice Conversion: State-of-the-Art and Future Work," presented at DAGA 2005, 31st German Annual Conference on Acoustics, Munich, Germany, 2005.

[13] O. Turk and L. M. Arslan, "Subband based voice conversion," presented at ICSLP '02, Denver, Colorado, USA, 2002.

[14] D. H. Klatt and L. C. Klatt, "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.

[15] D. Wong, J. Markel, and A. Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 27, pp. 350-355, 1979.

[16] H.-L. Lu, "Toward A High-Quality Singing Synthesizer with Vocal Texture Control," in *Department of Electrical Engineering*, vol. Doctor of Philosophy. San Francisco: Stanford University, 2002, pp. 120.

[17] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, pp. 505-519, 1995.