Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2008, Article ID 597971, 6 pages
doi:10.1155/2008/597971

Research Article

Developing a Child Friendly Text-to-Speech System

Agnes Jacob and P. Mythili

Division of Electronics, School of Engineering, Cochin University of Science and Technology, Kochi 682022, Kerala, India

Correspondence should be addressed to Agnes Jacob, nirag_2007@rediffmail.com

Received 11 November 2007; Revised 5 June 2008; Accepted 28 August 2008

Recommended by Owen Noel Newton Fernando

This paper discusses the implementation details of a child friendly, good quality, English text-to-speech (TTS) system that is phoneme-based, concatenative, easy to set up and use with little memory. Direct waveform concatenation and linear prediction coding (LPC) are used. Most existing TTS systems are unit-selection based, which use standard speech databases available in neutral adult voices. Here reduced memory is achieved by the concatenation of phonemes and by replacing phonetic wave files with their LPC coefficients. Linguistic analysis was used to reduce the algorithmic complexity instead of signal processing techniques. Sufficient degree of customization and generalization catering to the needs of the child user had been included through the provision for vocabulary and voice selection to suit the requisites of the child. Prosody had also been incorporated. This inexpensive TTS system was implemented in MATLAB, with the synthesis presented by means of a graphical user interface (GUI), thus making it child friendly. This can be used not only as an interesting language learning aid for the normal child but it also serves as a speech aid to the vocally disabled child. The quality of the synthesized speech was evaluated using the mean opinion score (MOS).

1. Introduction

There are various critical factors to be considered while designing a TTS system that will produce intelligible speech. Any TTS should be appealing to the child user whether it is used as a language learning aid or as a vocal aid. Children find learning more fun when their typed inputs are mapped to vocalized outputs. The first crucial step in the design of any concatenative TTS system is to select the most appropriate units or segments that result in smooth concatenation. This involves a tradeoff between longer and shorter units. In the case of shorter units, such as the phonemes, less memory is required. But, the sample collection and labeling procedures become more complex. The number of segments and the time required to cover the language increase steadily from word to the phoneme. In addition to being part of computationally manageable inventory of items, the synthesis segments chosen should capture all the transient and transitional information. The latter had been emphasized throughout this work, which in turn contributed to the smooth concatenation of speech segments in this TTS.

Even though speech is analog, phonemes are discrete. Inclusive of allophones, phonemes are less than hundred and are mainly the vowels, diphthongs, and consonants [1]. These allophones can be concatenated to produce smooth utterances without enormous computational effort compared to concatenation from the basic set of just 44 phonemes. Although phoneme appears to be an attractive linguistic unit for speech synthesis because of its limited number, most efforts [2] to string them together have failed. Pronunciation of phonemes depends on contextual effects, speaker's characteristics, and emotions. During continuous speech, the articulator movements depend on the preceding and the following phonemes. This causes some variations on how the individual phoneme is pronounced which lead to spontaneous variations in phoneme quality that is often known as coarticulation [3]. Thus, as per available facts [4], phoneme-sized building blocks were found to be unsatisfactory as synthesis segments because of the coarticulatory effects of the adjacent sounds. Further, one of the problems associated with segmenting words and storing the excised phonemes is the preservation of the characteristics of the sound, which is present at the transitions at the beginning

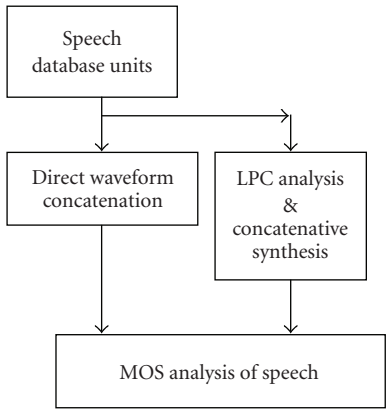

Figure 1: Overview of the TTS implementation.

and end of the segment. The characteristic sound of these transitions could be lost from both segments if they are smoothed together using signal processing techniques, resulting in a loss of naturalness in the utterance [5]. Hence, a major challenge in this method is that the boundaries between the phonemes correspond to areas that are acoustically volatile. Speech synthesized with phonemes as units is intelligible when each phoneme is represented by several allophones in the segment database. Different emotions and speaker characteristics could be implemented with such a database.

Figure 1 shows the major steps for implementing this work by the two methods. In the first method, a smooth, direct waveform concatenation of phoneme segments had been done, maximizing the speech quality in terms of naturalness and intelligibility. In the second method, LPC had been used to reduce memory requirements. A comparison of the above two methods, in terms of performance as well as the resources used, had been done using MOS. It was found that both methods gave a good quality TTS for children, which are not overloaded by too many analytical aspects and can be simulated in a short period compared to any other TTS. Sufficient care of the various linguistics aspects ensured natural sounding speech, inclusive of emotions.

2. Text/Vocabulary Selection

Selecting the vocabulary of the TTS is often referred to as fixing the target, since each of the utterances, for which this TTS system is designed, is called target and it is the speech output corresponding to a phonemic input. In the text selection phase, the target vocabulary as well as pitch, amplitude, and duration or speed of utterance was chosen. These varied styles correspond to different emotions. Being an experimental work with phonemes, the target was chosen to be a limited vocabulary of 635 different words covered by thirty-eight different phonemes and their allophones. Restriction of the targets words to a specific domain was seen to give better performance, since ultimately, the prosody is limited to the extent embedded in the recorded speech from which the segments are excised. Moreover, limiting the variety of utterances to one's anticipated needs can help one to provide sufficient samples of the segments without unnecessarily increasing the memory requirements of the database. This TTS had been initially designed to meet the minimal vocal requirements of a vocally impaired child. Alternately, it can also be modified to suit the language learning requirements of a normal child.

3. Text-To-Phoneme Conversion

The input text should be valid in the sense that it should belong to the set of target words. An error message is displayed in the graphical user interface (GUI), for cases falling outside the predetermined vocabulary (set of fixed inputs for which the speech database is designed). This project uses an event-driven front-end, which is simulated by means of a MATLAB program to transcribe input target words (the decided vocabulary) into their corresponding different phonemic forms. The international phonetic transcription (IPA) symbols are used. A dictionary-based approach has been used here for text-to-phoneme conversion. The pronunciation dictionary is initially designed in the context of the 635 different, typed input words only. The algorithmic complexity in the TTS design is considerably reduced due to this dictionary-based approach. The program stores words along with their correct pronunciation. This method is accurate and quick. The transcriptions are obtained from the sixteenth edition [6] of the *English Pronouncing Dictionary*. The pronouncing dictionary of this TTS provides essential information such as pronunciation of proper names and variant pronunciation than is usual in a general dictionary. Therefore, the primary aim of this TTS dictionary is to list pronunciation likely to be used by learners such as one that reflects the regional accent.

After considerable linguistics study and based on the findings from the literature survey, certain words are selected for the text corpus (to be recorded). These words are constituted by the phonemes of desired interest (present in the target words). Choosing specimen words for recording, followed by segmenting of these word recordings to extract the desired phonemes is the next crucial task and is related to linguistics.

4. Design of the Text Corpus

Careful design of the text corpus is an essential prerequisite to speech database design and involves the preparation of an inventory of words, which have the same phonemic constituents (along with contexts) of the target words. Care is taken to include three to five examples of words with the same allophone and these words are recorded. Even though the simplest approach is to add data to the speech database till the quality does not improve anymore, in all practical cases, there is a tradeoff between the quality and quantity. Each of the constituent phoneme segments of the target words is examined, and care is taken to select and record words having these very features. Examples of certain "design rules" formulated on the basis of the linguistics study are

Table 1: The design process.

Word (target)	Transcription	Words to record	Description
Eye	aI	Buy, why, bide B | aI, w | aI, b|aI|d	Diphthong. greater duration on first sound a: cases favored are word final or followed by voiced consonant.

given below [7]. These rules are also followed when selecting segments for concatenation as follows.

(1) A vowel preceding a voiced consonant in the same syllable tends to be longer than the same vowel preceding a voiceless consonant.

(2) The length of long vowels and diphthongs is very much reduced when they occur in syllables closed by consonants such as /p, t, k, s, h/.

(3) The consonant letter y can act as a vowel and as a consonant.

(4) The length of a phoneme is the least when it is in the middle of the word and maximum at the end.

The use of stressed words served to increase the duration of the phonemes by more than 20%. Since content words (words that are important for their meaning, e.g., nouns, adjectives, adverbs) are stressed, the text corpus to be recorded is constituted mostly by content words.

Based on the above criteria, a table was prepared with each row containing a word to be synthesized, its phonemic constituents, and examples of words with these constituent phonemes in varied positions in the word. A sample is shown in Table 1 above. The database is designed to offer sufficient coverage of the units to make sure that an arbitrary input sentence can be synthesized with more or less homogeneous quality.

5. Phoneme Database Development

A series of preliminary stages have to be fulfilled before the synthesizer can produce its first utterance. The database of WAV files is obtained by recording the natural voicing of the targets. A sampling frequency of 22.05 kHz was used to make the synthesized voice sound more pleasant. An amplitude resolution of 16 bits was used [8]. The recordings were done in male and female voices. Repeated segmentation of these speech files was done to excise phonemes/allophones from the speech database. As the output quality of any concatenative speech synthesizer relies heavily on the accuracy of segment boundaries in the speech database [9], manual method of segmentation was used.

In this work, the coarticulatory effect was put to good use by excising phonemes from different environments (surrounding phonemes), adding to the variability and naturalness of the database. These allophone segments were also stored as WAV files after appropriate labelling.

6. Concatenation

Relevant literature cites that concatenating and modifying the prosody of speech units without introducing audible artifacts are difficult [3]. In this work, this problem was overcame by appropriate linguistic design of the text corpus and careful preparation of the speech database. Moreover, the acoustic inventory used consists of a rich storage of needed allophones rather than phonemes. Several linguistic rules have been closely followed. For instance, we apply that vowels are key components [2] determining the synthesized voice quality. After manually editing the WAV files and trying out the direct waveform concatenation to identify the right constituent segments for any word in the predetermined vocabulary, the appropriate segments are concatenated programmatically to yield the synthesized speech. Sentences could also be synthesized with the prosody corresponding to those embedded in the segments [10]. Sentences made from segments of longer duration give rise to slow utterances and correspond to sad emotions, while sentences from short duration segments give rise to fast utterances corresponding to any happy, energetic person. These varied styles could be chosen using tags in the text entry. Allowing the choice between male and female voices provided an additional degree of customization [11]. A female voice, sampled at 22.05 kHz, was played back at a reduced sampling frequency of 17.5 kHz in order to produce a distinct male voice. Such voice conversion attempts proved to be amusing to children, instilled in them curiosity about the mechanism of speech production and improved their intellectual ability. This TTS system is a brilliant way to expose interested students to the basics of phonetics and motivates them to setup any TTS of their choice.

6.1. Parametric Representation of Speech

The second method uses the linear predictive coded (LPC) speech [3], wherein wave files are replaced by parametric models. In the LPC method of resynthesis, the voicing, pitch, gain, and LPC parameters were found for the down-sampled versions of the above constituent wave files, and speech was resynthesized. The source filter model [12] of speech production used here hypothesizes that an acoustic speech signal can be seen as a source signal (the glottal source, or noise generated at a constriction in the vocal tract), filtered with the resonances in the cavities of the vocal tract, downstream from the glottis, or the constriction. The LPC model was used for the prosody modification as it explicitly separates the pitch of a signal from its spectral envelope. Speech is parameterized by an amplitude control, voiced/voiceless flag, Fundamental Frequency (F0), and filter coefficients at a small interval. The F0 is the physical aspect of speech corresponding to perceived pitch. As a periodic signal, voiced speech has spectra consisting of harmonics of the fundamental frequency of vocal fold vibration. The loudness control is determined from the power of speech at the time frame of analysis. The concatenative approach to speech synthesis requires that speech samples should be stored in some parametric representation that will be

suitable for connecting the segments and changing the signal characteristics like loudness and F0. F0 extraction algorithms determine the voicing of speech as well as the fundamental frequency.

The LPC analysis is done for the same candidate WAV files used in the direct waveform concatenation. The pitch detection function (program) developed in MATLAB takes the speech audio signal and divides it into 30 milliseconds frames, over which speech is quasistationary. These overlapping frames start every 15 milliseconds and were further Hamming windowed to avoid distortion [3]. The function returns the pitch value in hertz for voiced frames, whereas it returns a zero for unvoiced frames. Monotone speech was produced from the synthesizer by replacing the pitch signal calculated by the function with a vector of constant values. The constant values, that were selected, were 100 Hz and 380 Hz, which correspond to male and child voices, respectively.

The Synthesis function returns the reconstructed audio. Using the voicing, pitch, gain, and LPC coefficients, each frame was synthesized. These were then put together to form the synthesized speech signal. The initial 30 milliseconds signal was created based on the pitch information. If the pitch is zero, the frame is unvoiced. This means that the 30 milliseconds signal needs to be composed of white noise. White noise is noise with a flat spectrum (uniform power spectral density) over the entire frequency range of interest. The term "white" is used in analogy with white light, which is a superposition of all visible spectral components.

Unvoiced excitation is usually modelled as such a white noise limited to the bandwidth of speech [3]. In this implementation, MATLAB function "randn" was used for producing white noise. For nonzero pitch, a 30 milliseconds signal was created with pulses at the pitch frequency. These initial signals were filtered using the gain and filter coefficients and then connected together in overlapping frames, for smooth transition from one frame to the next.

6.2. MOS Evaluation

Evaluating synthetic speech formally is difficult as there are many complex factors, dealing with intelligibility, naturalness, and the flexibility to simulate different voices and speaking rates. Due to lack of suitable standards for comparison, objective methods could not be used in this work. Hence, evaluating synthetic speech output was almost exclusively a subjective process. Certain subjective tests such as the dynamic rhyme test (DRT) are not realistic for practical application [3]. Therefore, mean opinion score (MOS) [13] has been used to evaluate the quality of this TTS, mainly in terms of intelligibility and naturalness. We have used the five level scales given in Table 2 as they are easy and provide some instant, explicit information.

An MOS rating greater than 4 indicates good quality. Any rating between 3.5 and 4 indicates that the utterance possesses telephonic communication quality. Ten volunteers without any known hearing disabilities participated in the MOS evaluation of the outputs of both phases. The listeners were all nonnative speakers of English. As is required, none were experts in TTS. There were five teachers (one of them well versed in linguistics), two 8-year-old kids (who are used to synthetic voices), two doctors, and 1 person who had recovered from a voice loss recently. The participants were briefed about the project.

Table 2: Scales used in MOS.

Rating	MOS
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Table 3: Sample MOS for sentences.

No.	Test sentences	MOS rating
1	Where are you going?	4.5
2	Please leave me alone.	4

Table 4: MOS for WAV concatenation.

No.	Words	MOS
1	Fine	5
2	Bill	4.9
3	Queue	4

7. Results and Discussion

Synthesis of polysyllabic words was done without any difficulty. Each participant randomly selected and listened to 25 words. The average MOS rating for each stimulus (TTS utterance) was later calculated and tabulated. As these tests were aimed at assessing the segmental as well as the suprasegmental characteristics, [14] the volunteers took further listening tests of a prepared list of 20 sentences and were asked to rate these using MOS. A sample is given in Table 3. The MOS tests were conducted individually for each listener. After hearing a test stimulus, the listener indicated his/her rating on a 5-point scale. The tests for both methods were administered at a stretch.

A sample of the MOS [13] evaluation of direct waveform concatenation of phonemes is as given below in Table 4. The average MOS ratings indicated that the speech output of the TTS is of good/"toll" quality.

However, as expected, results of the MOS evaluation of the LPC-based, parametric, concatenative TTS indicate that no words are rated as excellent. Listeners unanimously stated that the intelligibility of words increased considerably when these were embedded in sentence utterances rather than in isolation.

The memory comparison given in Table 5 highlights the advantage of incorporating the LPC model. There is considerable saving of memory though at the expense of the quality of the synthesized speech. The average memory gain factor was found to be 8.96, thus justifying the use of LPC as a means to achieve database compression. This is a significant

Table 5: Sample comparison of memory for LPC and WAV file storage.

Word	Constituent files	(X) WAV file storage (bytes)	(Y) LPC parametric storage (bytes)	(X)/(Y) gain factor
SO	S01SJ8 O1O8	24000	2760	8.7

achievement compared to the vast memory requirements of conventional unit selection-based methods.

Additionally, it has been found that the target vocabulary could be generalized in the sense that using the database of phonemes suitable to produce the predefined vocabulary, many other words could also be produced by suitable concatenation. Hence, this implementation is efficient.

As such, it can be implemented by any person with a basic knowledge in linguistics and programming. Since all can use the same basic program, this can, therefore, be self-implemented by students with minimal guidance from a tutor. Since children are often more receptive to certain voices like those of their teacher or parent, the database can be recorded in any of their preferred voices for better and enjoyable learning. The vocabulary also can be chosen as per the preferences of the child. This TTS in one's own voice or in any other preferred voice motivates the child learner to try out new words. The provision to record user's own voice and compare it with the TTS utterance provides sufficient motivation for the child learner to expand his vocabulary; if one keeps track of correct utterances of TTS and assigns appropriate scores to the user. Thus word building exercises can be made more interactive and amusing, as only correct words will be vocalized, whereby children can feel out the word and get immense satisfaction with each complete utterance.

Unlike unit selection concatenative systems [15] which make use of varied speech units and mostly engineering techniques like cost optimization and signal processing, this TTS implementation minimized its algorithmic complexity primarily by incorporating appropriate linguistic aspects like coarticulation and a carefully designed database to ensure smooth concatenation. Further reduction in algorithmic complexity was achieved by using table lookup methods for the grapheme to phoneme conversion. Though there are various speech aids for the vocally handicapped, any person who once possessed the ability to speak would normally prefer to use his/her own voice compared to any other robotic voice. Hence, this will also help children facing the risk of an impending vocal impairment due to some illness. Besides, this sort of speech synthesizer requires no additional expense for a person with a good computer along with speakers. With the proliferation of laptops and notebook computers, mobility is also not a problem.

8. Conclusion

In this paper, the implementation details of a child friendly phoneme-based concatenative TTS, with sufficient degree of customization and which uses linguistic analysis to circumvent most of the problems of existing concatenative systems, have been presented. The use of a dictionary-based approach for text-to-phoneme conversion along with a tailored speech database helps to avoid all algorithmic complexities and concatenation mismatches, characteristic of existing TTS systems. While conventional unit selection-based TTS requires hundreds of megabytes of memory, this TTS required only hundreds of kilobytes of memory.

Voice conversion feature has been incorporated in this TTS using the LPC method with provision for varying the voice quality over a wide range by varying the F0 values in the synthesis stage. Another feature of this work is that it was implemented using female voice, whereas most of the successful LPC-based TTSs have been implemented in male voice. This TTS further has add-on facility in that new words can be synthesized, after adding these words, their transcription, and constituent wave file names to their respective databases. The prosody of the utterance can be designed to vary depending on the nature of the recordings in the speech database from which the phoneme segments are excised. Thus, a simple, flexible, and efficient TTS that can be user defined has been setup with minimum resources to serve multiple purposes. Though this had been developed for English, it can be suitably modified for any other language. This TTS was found to be a successful vocal aid/language learning aid as the users were able to get a real feel of phonemes, the most basic speech units. The learning environment can be conditioned to any particular accent by using an appropriate combination of database and pronunciation dictionary. Alternately, content specific learning too can be encouraged implicitly. By suitable design of the TTS vocabulary and database, a child can be familiarized with all common terms associated with any specific topic. Thus, such a TTS helps the child user get acquainted with the regular as well as any other selective vocabulary.

References

[1] T. Parsons, *Voice and Speech Processing*, McGraw-Hill, New York, NY, USA, 1987.

[2] E. Keller, Ed., *Fundamentals of Speech Synthesis and Speech Recognition*, John Wiley & Sons, New York, NY, USA, 1994.

[3] D. O'Shaughnessy, *Speech Communications: Human and Machine*, Cambridge University Press, Cambridge, UK, 2001.

[4] C. Rowden, *Speech Processing*, McGraw-Hill, New York, NY, USA, 1992.

[5] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1976.

[6] D. Jones, *Cambridge English Pronouncing Dictionary*, Cambridge University Press, Cambridge, UK, 2003.

[7] T. Balasubramanian, *A Textbook of English Phonetics for Indian Students*, Macmillan India, New Delhi, India, 2003.

[8] VoiceSynthesis, http://www.hitl.washington.edu/scivw/EVE/.

[9] M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," *IEEE Communications Magazine*, pp. 99–104, 2002.

[10] M. Tatham and E. Lewis, "Improving text-to-speech synthesis," *Proceedings of the Institute of Acoustics*, vol. 18, no. 9, pp. 35–42, 1996.

[11] A. S. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," http://www.festvox.org/festival/.

[12] A. M. Kondoz, *Digital Speech Coding for Low Bit Rate Communication*, John Willey & Sons, New York, NY, USA, 1994.

[13] C. Delogu, A. Paoloni, and P. Pocci, "New directions in the evaluation of voice input/output systems," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, pp. 566–573, 1991.

[14] Y. Sagisaka, "Speech synthesis from text," *IEEE Communications Magazine*, vol. 28, no. 1, pp. 35–41, 1990.

[15] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. 465–468, Orlando, Fla, USA, May 2002.