Natural Speech Synthesizer for Blind Persons Using Hybrid Approach

Mukta Gahlawat\(^a,b\)*, Amita Malik\(^a\), Poonam Bansal\(^b\)

\(^a\) DeenBandhu ChotuRam University of Science & Technology Murthal, India.
\(^b\) Maharaja Surajmal Institute of Technology, Jankpuri, New Delhi, India

Abstract

The major challenges faced by the researchers in speech synthesis are intelligibility and naturalness. Intelligibility means easily understandable and naturalness means the quality of speech being very near to human speech. Due to dynamic nature of human speech it is very difficult to mimic it, as the same content of speech in different situations is having different prosodic parameters. This paper discusses an approach to develop a natural sounding speech synthesizer. The developed Text To Speech system was tested on blind persons using subjective listening test. Test was performed using mean average score (MOS) and it was done on ten blind persons of age group varies from 14 years to 42 years. Five parameters naturalness, intelligibility, usability, localization awareness, expressions were considered for analysis of the speech synthesizer. As a result, good MOS was received for naturalness and usability, fair MOS for intelligibility and localization.

Keywords: Speech, Text to Speech, Expressive Speech, Unit selection, Concatenative Speech Synthesis

1 Introduction

Speech is the most natural way to communication between two or more persons. For effective communication expressions, clarity of speech and pronunciation play an important role to deliver the message correctly. When the speech synthesizer is developed, the researcher always tries to synthesize the speech as close as possible to human speech. Different peoples have different characteristics like pitch, prosody, accent, pronunciation etc. so it is very difficult to follow the standard speech characteristics all over the world. Even the individual’s speech is full of variations depending upon his mood, biological fitness, and different state of mind. These are some reasons that justify why the
natural sounding speech is still a state of art after having a long history of research. Speech synthesis means conversion of written text into spoken words by concatenating speech waveforms. There are number of ways of speech synthesis as discussed by (Lemmetty, 1999) in his review. First way, is the articulatory synthesis where the human vocal organs and articulation processes are modeled. Speech is created by digitally simulating the flow of air through the representation of the vocal tract. It produces high-quality synthetic speech but this technique is very hard to implement. Second technique is the Formant speech synthesis that involves an acoustic model for generating synthesized speech output. It does not use human speech samples instead there are a number of parameters which needs to be considered like fundamental frequency, voicing, and noise levels etc. This technique lacks naturalness of speech. Third method is the conatenative synthesis of speech which is considered as best for natural sounding speech synthesis because it is based on the concatenation of pre recorded segments of speech. Waveform is generated by selecting and concatenating the appropriate units from a database consisting of different types of speech units (like phones, diphones, syllables, words, phrases). Other methods like HMM based and linear prediction methods also exist in literature.

The aim of this work is to generate natural sounding speech; hence concatenative speech synthesis is implemented using unit selection algorithm (A. Hunt, 1996) (Black, 2003). For developing natural speech, a hybrid approach where the expressions and spatial parameters are unified, is used to make synthesized speech more natural. The normal vision persons can easily understand the expression of the speaker just by seeing his facial gestures but for visually impaired person it is not possible to indentify the mood or expressions of speaker. Moreover, majority of Text To Speech Synthesizer (or TTS) software’s that are used by blind persons lack naturalness and expressions. Additionally, during testing one interesting input from listeners was received that this TTS system has the personalized database recorded by non-native speaker of English so they were able to understand the word more easily as compare to the software they were using in their labs. They mentioned the reason that the accent and pronunciation of words are same as they speak. The approach of adding expressions with spatial speech is purposed .This paper is organized in 6 sections, section 2 describe the related work, section 3 gives details of proposed approach , section 4 includes testing details followed by results obtained and last section include conclusion and future scope.

2 Related Work

The Speech synthesis is not new branch of research, it has a long history. Generating natural sounding speech is a big challenge of this field. When we talk about emotional speech, there are many authors who have done emotional speech synthesis using various techniques and in various emotions. (Akemi Iida, 2003) Synthesize the emotional speech by a corpus-based concatenative speech synthesis system using large emotional speech corpora. They have considered three kinds of emotions anger, joy, and sadness. They have created the corpora for Japanese language. (Daniel Erro, 2010) Designed the system which perform emotion conversion by manipulating prosody. Intonation, duration and intensity were taken as three prosody parameters. (Aimilios Chalamandaris, August 2010) implemented the unit selection technology into screen reading environments. They carried out subjective test using MOS to evaluate the resulting system. (Haojie Zhang, 2012) Synthesize the emotional speech by adjusting fundamental frequency and formant transition. (Roberto Barra-Chicote, 2010) have generated emotional speech by integrating unit selection and HMM based synthesis and found that unit selection require improvement in prosodic modeling and HMM require improvement in spectral modeling. Also there were some emotions which were not reproduced by either method. (Tonnesen & Steinmetz, 1993) Had work on synthesis of 3D speech. They described various ways to generate 3D sound, challenges for spatial sound and its applications. (Jaka Sodnik, 2011) Designed multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. They describe various benefits and drawbacks of simultaneous spatial sounds in auditory interfaces for visually impaired and blind computer users. They took two different auditory interfaces in spatial and non-spatial condition to represent the hierarchical menu structure of a simple word processing.
application. Their hypothesis was that using multiple spatial sounds simultaneously will be faster and more efficient than non spatial. But after testing on blind people they found that multiple simultaneous sounds requires the entire capacity of the auditory channel and total concentration of the listener and performance was slow. (Tomažič, 2009) Also worked on spatial speaker using 3Dimensional Java Text To Speech conversion.

3 Natural Speech Synthesizer

The approach for generating natural sounding speech synthesizer is described in this section. Firstly, an emotional corpus was created for three different emotions-neutral, happy and sad. The database is recorded with the help of one female speaker. After recording the segmentation of the database was done. The user inputs the text and proper units were selected from the database (Mukta Gahlawat, 2013). Speech synthesis was performed using TTSBOX (Thierry Dutoit, 2005). An audio speech signal was generated that is then converted to spatial speech and audio output is generated. After the audio output, the spatial speech was generated to give spatial effect. Spatial sound (Jaka Sodnikn, 2011) is the sound that we hear in everyday life. Sounds come at us from all directions and distances (Tonnesen & Steinmetz, 1993). The brain gets cues about the direction and distance of objects from us from the surrounding environment. Spatial sound gives a sense of the sound's position as recorded by the microphones. The human head filters the incoming sounds. For developing the spatial speech synthesizer, we have used Head-Related Transfer Functions (HRTF) and Open AL audio libraries. HRTF (Corey I. Cheng, 2001) (Kulkarni, 1995) is a response that characterizes how an ear receives a sound from a point in space. Open Audio Libraries (openAl.) is an audio library that contains functions for playing back sounds and music in a game environment. It helps the programmer to load the sound and control certain characteristics such as position, velocity, direction and angles that determine how the sound is traveling. All sounds are positioned relative to the listener which represents the current place where the user is.

3.1 Database Design

The database was created using open source software’s and the details of the process are described in Expressive speech synthesizer (Mukta Gahlawat, 2013). Using the same approach the set of new database was created. This unit of recording was sentences. The language used is English in Indian accent. The database consists of around 849 words in all three emotions. Among these 525 were distinct words in 168 sentences. There were around 324 words which are present in database more than once. Table 1 summarizes the database.

Table 1: Summary of Database used

| Number of sentences | Number of distinct words | Total Number of words in each Emotion |
|---------------------|--------------------------|--------------------------------------|
| 56                  | 175                      | 283                                  |

| Number of sentences | Number of distinct words | Total Number of words in each Emotion |
|---------------------|--------------------------|--------------------------------------|
| 56*3=168            | 175*3=525                | 283*3=849                            |
4 Testing

The intention of this work is to build the Natural Sounding Speech Synthesizer (or NSS) by adding single spatial sound to the Expressive Speech Synthesizer (ESS). To test the quality of speech, testing was done on blind persons. The NSS was tested on five parameters. Testing was done on 10 persons (9 were blind, 1 was partially blind). Among these, 7 blind students, 3 bind teachers. The minimum age was 14 and maximum age was 42 years. The average age of our test subject comes was 19.4. There were 7 males and 3 females. Before doing actual testing the listeners were familiarized with NSS. Testing was done at at Akhil Bhartiya Netrahin Sangh, Residential School and Training Center for Blinds Raghubir Nagar, New Delhi. Laptop and headphones were used in their computer labs. The blind students were using JAWS in their lab for doing their work. One by one students were called in computer laboratory and using our headphones testing was done. Individual feedback was taken from them. Testing of NSS was performed at two level word levels and at sentence level. Six testing words and six sentences were taken for performing test on each individual. This synthesizer can only synthesize the words that are present in database. Five parameters were Naturalness, Intelligibility, Directional Awareness, Expressiveness and over all usability. For scoring Mean Opinion Score (MOS) (Deller, 1993) was used. Each listener was asked to provide the score under each parameter with 0 and 5 as minimum and maximum scores respectively. 0 means unsatisfactory and 5 means excellent rating. Table 2 gives the details of Mean Opinion Score.

Table 1: Meaning of MOS

| Mean Opinion Score (MOS) | Quality       |
|--------------------------|---------------|
| 5                        | Excellent     |
| 4                        | Very Good     |
| 3                        | Good          |
| 2                        | Fair          |
| 1                        | Poor          |
| 0                        | Unsatisfactory|

5 Result and Discussion

After performing subjective listening test satisfactory results were obtained for the subject. The result shows that single spatial sound if integrated with expressive speech make speech more natural and interesting. On the basis of input received from listeners, the graphs were plotted which are shown in figures 1-5 below. The first parameter to test the NSS was Naturalness. It means how much the speech resembles to human voice. Figure 1 shows MOS for Naturalness. The average score was 4.6, which shows the speech of NSS was very natural. Second parameter was intelligibility, which means how many words are recognized correctly. Figure 2 shows MOS for intelligibility, the average MOS for intelligibility was 4.4. Third parameter was the Directional Awareness that signifies how many directions were recognized when speech was played. Three directions were recognized left, right and center. Figure 3 shows mean opinion score for Directional Awareness and the average MOS for direction identification was 4.6. Fourth parameter was Expression, which were used to predict the mood of speaker. Figure 4 shows mean opinion score for emotion recognition and the average MOS for emotion recognition was 4 which were least among all parameters. Fifth parameter was over all usability, in which listeners were asked if how much the NSS was useful for them. Figure 5 shows mean opinion score for overall usability of NSS, and listeners give average 4.7 MOS for overall
usability which was highest among all the five parameters. In addition to mean opinion score, we have also asked to share their experience on such application. All the listeners give almost same experience. It was different and new experience for them which they had never felt before. They said that best part of Natural Speech Synthesizer was that it was very lively because of expressions. Secondly, the sentences and words were recorded in our accent i.e Indian accent, so they could easily understand the words as compare to the software which they were using. The results show that by adding the spatial parameters in expressive speech have made speech more natural as perceived by blind persons.

**6 Conclusion and Future Scope**

Adding spatial parameters to expressive speech synthesizer increase the naturalness and usability to a satisfactory level. The feedback received from the listeners shows that by adding spatial speech in expressive speech not only makes the speech natural but quite intelligible also. We can also use this hybrid concept for developing other applications for blind or visually impaired persons. Some of the suggested applications where this approach can be implemented are story telling for disabled persons.
Secondly, this approach can also be used for developing computer based games for disabled or blind persons. As far as future scope is concerned, further improvement can be done on the quality of synthesizer for expression. Additionally, database can be increased by adding some more emotions to it. Lastly, work can be done to add some more directions.

References

(n.d.). Retrieved February 5, 2013, from openAl.: http://www.openal-soft.org/

A. Hunt, A. B. (1996). Unit Selection in a concatenative speech synthesis using a large database. ICASSP, (pp. 373-376). Atlanta, Georgia.

Aimilios Chalamandaris, S. K. (August 2010). A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers. IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, pp-1890-189.

Akemi Iida, N. C. (2003). A corpus-based speech synthesis system with emotion. Elsevier journal of Speech Communication, vol 40 , pp- 161–187.

Black, A. (2003). Unit Selection and Emotional Speech. Eurospeech. Geneva, Switzerland.

Corey I. Cheng, A. S. (2001). Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space. J Audio Eng Soc , Vol. 49, No 4.

Daniel Erro, E. N. (2010). Emotion Conversion Based on Prosodic Unit Selection. IEEE Transaction on Audio, speech, and language processing , Vol. 18, No. 5. pp 974-983.

Deller, J. P. (1993). Discrete-Time Processing of Speech Signals. New york: Macmillan Publishing Company.

Haojie Zhang, Y. Y. (2012). Fundamental Frequency Adjustment and Formant Transition Based Emotional Speech Synthesis. 9th International Conference on Fuzzy Systems and Knowledge Discovery, (pp. 1797-1801).

Jaka Sodnikn, G. s. (2011). Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. Int. J. Human-Computer Studies , vol 69,100–112.

Kulkarni, A. (1995). On the Minimum-Phase Approximation of Head-Related Transfer Functions. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, (p. IEEE catalog no. 95TH8144.).

Lemmetty, S. Review of Speech Synthesis Technology. Master's thesis. Helsinki University, Finland.

Mukta Gahlawat, A. M. (2013). Expressive Speech Synthesis System Using Unit Selection. Mining Intelligence and Knowledge Exploration (pp. Volume 8284 , 391-401). Springer Lecture Notes in Computer Science.

Roberto Barra-Chicote, J. Y.-G. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Elevier journal of Speech Communication.

Thierry Dutoit, F. F. TTSBOX: A Matlab Toolbox for Teaching Text-to-Speech Synthesis. ICASSP. Philadelphia.

Tomazič, J. S. (2009). Spatial Speaker: 3D Java Text-to-Speech Converter. In Proceedings of the World Congress on Engineering and Computer Science., (p. Vol. II WCECS 2009). San Francisco, USA.

Tonnesen, C., & Steinmetz, J. (1993). 3-D sound Synthesis. Retrieved 2014, from Washington, The Encyclopedia of Virtual Environments: http://www.hitl.washington.edu/scivw/EVE/I.B.1.3DSoundSynthesis.html