

Bi-Lingual Text to Speech Synthesis System for Urdu and Sindhi

Azhar Ali Shah, Abdul Wahab Ansari and Lachhman Das
aas_lakyari@hotmail.com, whabanari@hotmail.com and lac_d@hotmail.com
Institute of IT, University of Sindh, Jamshoro, Pakistan

Abstract: Text to Speech Synthesis along with the Speech Recognition is widely used throughout the world to enhance the accessibility of the information and enable even the disabled persons to interact with the computers in order to get the potential benefit from this high-tech revolution. In this paper we introduce a bi-lingual novel algorithm for the synthesis of Urdu and Sindhi language text.

The devised bi-lingual algorithm uses knowledge based approach along with the hybrid rule based and concatenative acoustic methods to provide efficient and accurate conversion of Urdu and Sindhi text into the high quality speech. The algorithm has been implemented in the VB programming language with a GUI based interface. The proposed system works with high accuracy and has a great potential to be used for variety of applications. The system is versatile enough and can be used for speech recognition also.

Keywords: TTS, Articulatory, Formant, Concatenative, waveforms, phones, diaphones

1. INTRODUCTION

Text-to-speech (TTS) synthesis technology gives machines the ability to convert arbitrary text into audible speech, with the goal of being able to provide textual information to people via voice messages. Key target TTS applications in communications include: voice rendering of text-based messages such as email or fax as part of a unified messaging solution, as well as voice rendering of visual/text information such as web pages. In the more general case, TTS systems provide voice output for all kinds of information stored in databases such as phone numbers, addresses, navigation information, restaurant locations and menus, and movie guides. Ultimately, given an acceptable level of speech quality, TTS could also be used for reading books i.e., Talking Books.

1.1 Synthesis Methods

There exist several different methods to synthesize speech. Each method falls into one of the following categories: articulatory synthesis, formant synthesis, and concatenative synthesis

1.1.1 Articulatory synthesis

It uses computational biomechanical models of speech production, such as models for the glottis (that generates

the periodic and aspiration excitation) and the moving vocal tract. Ideally, an articulatory synthesizer would be controlled by simulated muscle actions of the articulators, such as the tongue, the lips, and the glottis. It would solve time-dependent, 3-dimensional differential equations to compute the synthetic speech output [1-3]. Unfortunately, besides having notoriously high computational requirements, articulatory synthesis also, at present, does not result in natural-sounding fluent speech.

1.1.2 Formant synthesis

It uses a set of rules for controlling a highly simplified source-filter model that assumes that the (glottal) source is completely independent from the filter [4]. The filter is determined by control parameters such as formant frequencies and bandwidths. Each formant is associated with a particular resonance of the vocal tract. The source generates either stylized glottal or other pulses for periodic sounds and noise. Formant synthesis generates highly intelligible, but not completely natural sounding speech. However, it has the advantage of a low memory footprint and only moderate computational requirements.

1.1.3 Concatenative synthesis

This type of synthesis uses actual snippets of recorded speech that were cut from recordings and stored in an inventory called "voice database", either as "waveforms" (uncoded), or encoded by a suitable speech coding method. Elementary "units" i.e., speech segments are, for example, phones (a vowel or a consonant), or phone-to-phone transitions ("diphones") that encompass the second half of one phone plus the first half of the next phone (e.g., a vowel-to-consonant transition). Concatenative synthesis itself then strings together (concatenates) units selected from the voice database, and, after optional decoding, outputs the resulting speech signal. Because concatenative systems use snippets of recorded speech, they have the highest potential for sounding "natural".

1.2 Urdu and Sindhi Language Alphabet and Phonemes

Both Urdu and Sindhi languages share the same type of script, that is, Perso-Arabic Script which is written from right to left. There are some peculiarities that can be seen in Perso-Arabic scripts. For example these scripts join letters with each other, and therefore letters have different forms as per their position in a ligature [5-6]. These different forms for a Sindhi character 'Bhay' are shown in figure 1 (a, b, c, and d).

4. IMPLEMENTATION OF THE TTS FOR URDU AND SINDHI

Based on the steps illustrated in Figure 3 the code has been written in MS Visual Basic to develop a GUI based user interface as shown in Figure 4. The interface provides a menu bar with different menus providing the options for selection of a particular language, keyboard and speaker. It also provides a bilingual text editor where the user can enter the Urdu or Sindhi text. Once the user enters the text and clicks on the speak button then the interface displays the windows media recorder with the synthesized speech corresponding to Urdu / Sindhi text loaded. The user can just play the sound or save it at any destination for later retrieval.



Figure 4. Illustration of the Bilingual TTS

5. RESULTS AND DISCUSSION

The implemented TTS system has been tested by writing different Urdu and Sindhi sentences in the bilingual text editor as shown in Figure 5 and Figure 6.



Figure 5. TTS Interface with Urdu text



Figure 6. TTS Interface with Sindhi text

As the user clicks on the speak button then the interface displays windows sound recorder with the corresponding sound loaded. This is shown in Figure 7. Now user can play the sound or save it at a particular destination.



Figure 7. Synthesized Urdu / Sindhi speech loaded into the windows sound recorder.

6. CONCLUSION

The proposed bilingual TTS system for Urdu and Sindhi has been implemented using the concatenative synthesis method. The system utilizes a common acoustic inventory for both the languages hence reducing the number of acoustic units and improving the overall efficiency of the system. The system provides high quality Urdu and Sindhi speech and can be further expanded to incorporate expressive and visual text-to-speech (VTTS) strategies in future.

ACKNOWLEDGEMENT

The authors are indebted to various individual Urdu/Sindhi linguistics for their valuable suggestions and guidance.

REFERENCES

- [1] Sondhi, M. M., and Schroeter, J., Speech Production Models and Their Digital

- Implementations, in: The Digital Signal Processing Handbook, V. K. Madisetti, D. B. Williams (Eds.), CRC Press, Boca Raton, Florida, pp. 44-1 to 44-21, 1997.
- [2] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System", IEEE Proceedings, August 2000.
- [3] Alan W Black and Kevin A. Lenzo, "Multilingual text-to-speech synthesis", EUROSPEECH '99, Budapest, Hungary, Sept. 1999
- [4] Matthias Jilka, Ann K. Syrdal, Alistair D. Conkie and David A. Kapilow, "Effects on TTS quality of methods of realizing natural prosodic variations", ICPhS 2003, BARCELONA, SPAIN, AUGUST 2003
- [5] Soomro H.K, Shah A.A & Shaikh, A.A.G, "*Development of Computerized Sindhi to English and English to Sindh Dictionary*", MUET Res. J. of Engineering & Technology [ISSN 0254-7821], Vol. 23, No.4, October, 2004, p 289-296.
- [6] Chowdhry B.S, Memon A.R, and Memon N.U, "Voice Digitization for Sindhi Speech Synthesis", MUET Res. J. of Engineering & Technology, Vol. 13, No.3, July, 1994, p 1-8.
- [7] N. Fatima and R. Aden, "Vowel Structure of Urdu", 2003. *CRULP Annual Student Report* published in *Akhbar-e-Urdu*, April-May, National Language Authority, Islamabad, Pakistan.