Speech Transaction for Blinds Using Speech-Text-Speech Conversions

Johnny Kanisha[1] and G. Balakrishanan[2]

[1] Research Scholar, Anna university, Trichirappalli
[2] Director, Indra Ganesan College of Engineering, Trichirappalli

Abstract. Effective human computer interaction requires speech recognition and voice response.In this paper we present a concatenative Speech-Text-Speech(STS) system and discuss the issues relevant to the development of perfect human-computer interaction.The new STS system allows the visually impaired people to interact with the computer by giving and getting voice commands.Audio samples are collected from the individuals and then transcribed to text.A text file is used ,where the meanings for the transcribed texts are stored.In the synthesis phase,the sentences taken from the text file are converted to speech using unit selection synthesis.The proposed method leads to a perfect human-computer interaction

Keywords: STS method, Speech Synthesis, Speech recognition.

1 Introduction

Speech is one of the most vital forms of communication in everyday life. On the contrary the dependence of human computer interaction on written text and images makes the use of computers impossible for visually and physically impaired and illiterate masses.Speech recognition and speech generation together can make a perfect human computer interaction. Automatic speech generation from natural language sentences can overcome these obstacles. In the present era of human computer interaction, the educationally under privileged and the rural communities of any country are being deprived of technologies that pervade the growing interconnected web of computers and communications. Although human computer interaction technology has improved significantly in recent decades, current interaction systems still output simply a stream of words in speech. In speech recognition the unannotated word stream lacks useful information about punctuation and disfluencies that could assist the human readability of speech transcripts [1]. Such information is also crucial to subsequent natural language processing techniques, which typically work on fluent and punctuated input.Speech recognition and speech synthesis as separate phases will not give a good human computer interaction. Recovering structural information in speech and synthesisng the words has thus become the goal of many studies in computational speech processing [2], [3], [4], [5], [6], [7], [8]. We describe our approach that comprises advanced method for speech recognition using discrete wavelet transform and unit selction method for speech synthesis[22].

N. Meghanathan et al. (Eds.): CCSIT 2011, Part I, CCIS 131, pp. 43–48, 2011.

2 Problem Statement

In this paper we explain the concatenation of speech recognition and speech generation which leads to a small dictionary system that can be used by visually impaired people. The proposed approach can be called as STS (Speech-Text-Speech)method of speech transaction where a human can dictate a word ,that can be saved in a text file using STT(Speech to Text)System and the related meaning , searched from another file can be given in speech by the computer using TTS(Text to Speech) method. Speech recognition can be done by using the advanced speech recognition method based upon discrete wavelet transform that has 98% of accuracy.

In this method, first, the recorded signal is preprocessed and denoised with Mels Frequency Cepstral Analysis.The feature extraction is done using discrete wavelet transform (DWT) coefficients; Then these features are fed to Multilayer Perceptron (MLP) network for classification. Finally, after training of neural network ,effective features are selected with UTA algorithm. The speech syhthesis can be done by using concatenative speech synthesis which depends upon the unit-selection method of speech synthesis.

3 Architecture of Speech Recognition Method

The architecture of our speech recognition system has been shown in the figure below. Our speech recognition process contains four main stages:

1. Acoustic processing that main task of this unit is filtering of the white noise from speech signals and consists of three parts, Fast Fourier Transform, Mels Scale Bank pass Filtering and Cepstral Analysis.
2. Feature extraction from wavelet transform coefficients.
3. Classification and recognition using backpropagation learning algorithm.
4. Feature selection using UTA algorithm [9].

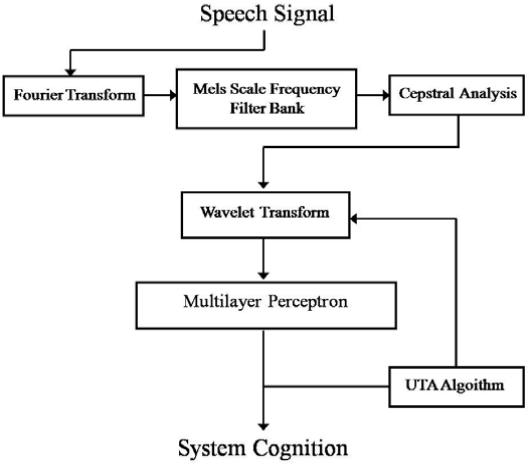

Fig. 1. Speech recognition

4 Unit Selection Synthesis

Unit selection synthesis uses large databases recorded speech[22]. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram[25]. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighbouring phones. Unit selection provides the greatest naturalness, because it applies only small amounts of digital signal processing (DSP) to the recorded speech.

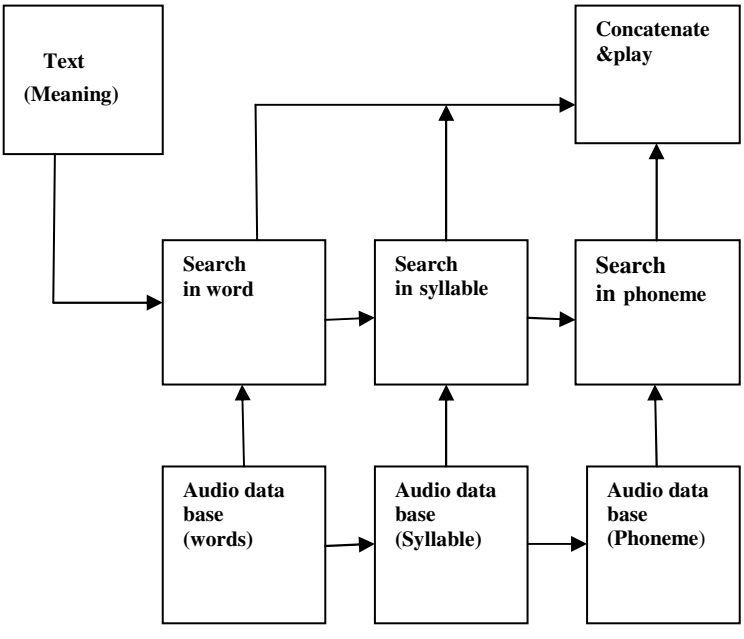

Fig. 2. Speech synthesis

5 Speech-Text–Speech Method

The Speech recognition method using discrete wavelet transform and the speech synthesis method using unit selection process are put together to form STS method.This concatenation is done to create a dictionary application for the blinds to speak a word to the system and to get the meaning for the transcribed word.

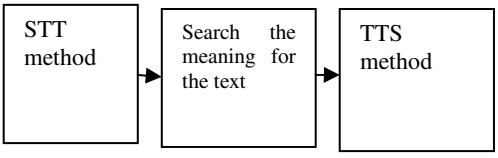

Fig. 3. The STS System

6 Steps for Dictionary System

1. The audio input given by the user should be transcribed to text using the speech recognition method and played back to the user using speech synthesis system

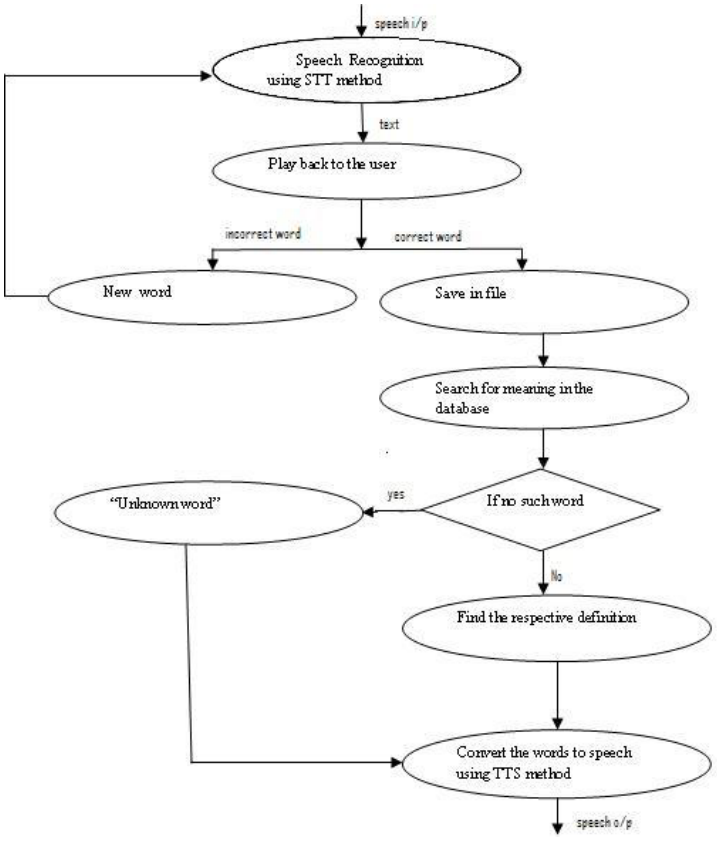

Fig. 4. STS Flow Chart

2. a.If the word is correct and accepted by the user ,the word will be saved in a text file
   b.If the word is incorrect then the user has to repronounce it within 10 seconds from the time he heard the word
3. In case of 2.a,the database is interfaced and the word will be checked with the words in the database and the respective meaning defined there will be identified
4. The identified text will be given back to the user in speech through speech synthesis system

7 Conclusion

In this paper, we discussed the issues relevant to the development of perfect human-computer interaction for the benefits of blinds. As man-machine interaction is an appreciated facility even outside the research world,especially for the persons with disabilities, we have designed the STS system,where the blinds can use the system as an audio dictionary.This can be extended to use the web dictionaries.Although being able to give and get speech make the STS system more flexible, the interaction quality may significantly decline unless the recognition and synthesis is done properly.The quality of interaction can be improved if the STT and the TTS system are provided with more descriptive prosody information.

References

[1] Jones, D., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D., Zissman, M.: Measuring the readability of automatic speech-to-text transcripts. In: Proc. of Eurospeech, pp. 1585–1588 (2003)
[2] Heeman, P., Allen, J.: Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. Computational Linguistics 25, 527–571 (1999)
[3] Kim, J., Woodland, P.C.: The use of prosody in a combined system for punctuation generation and speech recognition. In: Proc. of Eurospeech, pp. 2757–2760 (2001); peech transcripts, In: Proc. of ISCA Workshop: Automatic speech Recognition: Challenges for the Millennium ASR-2000, pp. 228-235 (2000)
[4] Gotoh, Y., Renals, S.: Sentence boundary detection in broadcast
[5] Kompe, R.: Prosody in Speech Understanding System. Springer, Heidelberg (1996)
[6] Snover, M., Dorr, B., Schwartz, R.: A lexically-driven algorithm for disfl uency
[7] Kim, J.: Automatic detection of sentence boundaries, disfluencies, and conversational fillers in spontaneous speech. Master's thesis, University of Washington (2004)
[8] Johnson, M., Charniak, E.: A TAG- based noisy channel model of speech repairs. In: Proc. of ACL (2004)
[9] Meysam, M., Fardad, F.: An advanced method for speech recognition. World Academy of Science,Engineering and Technology (2009)
[10] Kirschning. Continuous Speech Recognition Using the Time-Sliced Paradigm. MEng.Dissertation, University Of Tokushinia (1998)
[11] Tebelskis, J.: Speech Recognition Using Neural Networks, PhD. Dissertation, School Of ComputerScience, Carnegie Mellon University (1995)

[12] Tchorz, J., Kollmeier, B.: A Psychoacoustical Model of the Auditory Periphery as Front-endforASR. In: ASAEAAiDEGA Joint Meeting on Acoustics, Berlin (March 1999)
[13] Clark, C.L.: Labview Digital Signal Processing and Digital Communications. McGraw-Hill Companies, New York (2005)
[14] Kehtarnavaz, N., Kim, N.: Digital Signal Processing System-Level Design Using Lab View. University of Texas, Dallas (2005)
[15] Kantardzic, M.: Data Mining Concepts, Models, Methods, and Algorithms. IEEE, Piscataway (2003)
[16] Lippmann, R.P.: An Introduction to Computing with neural nets. IEEE ASSP Mag. 4 (1997)
[17] Martin, H.B.D., Hagan, T., Beale, M.: Neural Network Design. PWS Publishing Company, Boston (1996)
[18] Dietterich, T.G.: Machine learning for sequential data: A review. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 15–30. Springer, Heidelberg (2002)
[19] MathWorks. Neural Network Toolbox User's Guide (2004)
[20] Kishore, S.P., Black, A.W., Kumar, R., Sangal, R.: Experiments with unit selection Speech Databases for Indian Languages
[21] Sen, A.: Speech Synthesis in India. IETE Technical Review 24, 343–350 (2007)
[22] Kishore, S.P., Black, A.W.: Unit size in Unit selection Speech Synthesis. In: Proceedings of Eurospeech, Geneva Switzerland (2003)
[23] Kishore, S.P., Kumar, R., Sangal, R.: A data – driven synthesis approach for Indian Languages using syllable as basic unit. In: Proceedings of International Conference on National Language Processing, ICON (2002)
[24] Kawachale, S.P., Chitode, J.S.: An Optimized Soft Cutting Approach to Derive Syllables from Words in Text to Speech Synthesizer. In: Proceedings Signal and Image Processing, p. 534 (2006)
[25] Segi, H., Takagi, T., Ito, T.: A Concatenative Speech Synthesis Method using Context Dependent Phoneme Sequences with variable length as a Search Units. In: Fifth ISCA Speech Synthesis Workshop, Pittsburgh
[26] Lewis, E., Tatham, M.: Word and Syllable Concatenation in Text to Speech Synthesis
[27] Gros, J.Z., Zganec, M.: An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis