International Journal on Applications in Engineering and Technology
Volume 4: Issue 1: March 2018, pp 18 - 23. www.aetsjournal.com ISSN (Online) : 2455 - 0523

STATISTICAL PARAMETRIC SPEECH SYNTHESIS FOR WAVE FORM FRAME WORK REPRESENTATION USING BLIND PEOPLE

R.SUGANTHI , R.VETRIVENTHAN, SENTHIL KUMARAN

Abstract— A number of developing countries continue to provide educational services to students with disabilities in "segregated" schools. Also all students, regardless of their personal circumstances, have a right of access to and participation in the education system, according to their potential and ability. However, with the rapidly growing population and increasing number of people with blindness along with other disabilities, need for use of technology in the field of education has become imminent. With existing system of competitive examination, students face problems while interacting with the system, misunderstandings arising due to human mediator and also an ability to cope-up with the other students. Our project, through the use of speech technology, attempts to provide solutions for some of these issues by creating an interactive system. Thus, the application will help in creating an environment that provides equal opportunities for all the students in taking up competitive exams. This will improve the current educational system for blinds career.

Keywords— Speech recognition, blind people, android application.

I. INTRODUCTION

Blind people confront a number of visual challenges everyday – from reading the label on a frozen dinner to figuring out if they're at the right bus stop. There has been no proper website for providing online examination for blind people. By considering the above disadvantage we will be going to create an entire website for blind people to admit in online examination like a normal people without any difficulties. (Here, all outputs through in voice) Normally The Web Based Exam Management System has been developed to support automatic grading, exam archiving, and exam administration using the as a delivery vehicle. In most of them, the widely used questions are correspondence to Intended Learning Outcome (ILO) for the courses, and it should be easily judged and evaluated online by comparing with the correct answers. The typical questions include yes/no questions, multiple choice/ single answer questions, multiple-choice multiple-answer questions, matching questions, numeric questions, and essay questions. This system is built based on open source technology.Hence this application will allow visually impaired people to appear for test in more convenient and efficient way.

Accuracy of speech systems differ in vocabulary size and confusability, speaker dependence vs. independence, modality of speech (isolated, discontinuous, or continuous speech, read or spontaneous speech), task and language constraints. The basic element can be a phoneme for continuous speech or word for isolated words recognition. Dictionary is used to connect acoustic models with vocabulary words. Language model reduces the number of acceptable word combinations based on the rules of language and statistical information from different texts. Speech recognition systems, based on hidden Markov models are today most widely applied in modern technologies. They use the word or phoneme as a unit for modeling. The model output is hidden probabilistic functions of state and can't be deterministically specified. State sequence through model is not exactly known. Speech recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols.

II. STATISTICAL PARAMETRIC SYNTHESIS

A. Overview of a typical system:

A typical HMM-based speech synthesis system. It consists of training and synthesis parts. The training part is similar to those used in speech recognition systems. The main difference is that both spectrum (e.g., melcepstral coefficients and their dynamic features) and excitation parameters are extracted from a speech database and modeled by context-dependent HMMs (phonetic, linguistic, and prosodic contexts are taken into account). To model log F0 sequence which includes unvoiced regions properly, multi-space probability distributions are used for the state output stream for log F0. Each HMM has state duration densities to model the temporal structure of speech. As a result, the system models spectrum, excitation, and durations in a uniſed framework. The synthesis part does the inverse operation of speech recognition. First, an arbitrarily given text corresponding an utterance to be synthesized is converted to a context-dependent label sequence and then the utterance HMM is constructed by concatenating the contextdependent HMMs according to the label. Statistical parametric synthesis might be most simply

R.Suganthi , Master of Computer Applications , Meenaakshi Ramasamy Engineering College , Thathanur , Ariyalur Dt , Tamil Nadu .

R.Vetriventhan , BE , MTech , MISTE , Head of the department , Department of MCA , Meenaakshi Ramasamy Engineering College , Thathanur , Ariyalur Dt , Tamil Nadu .

Dr.Senthil Kumaran , ME Phd , Managing Director , Meenaakshi Ramasamy Engineering College , Thathanur , Ariyalur Dt , Tamil Nadu.

described as generating the average of some set of similarly sounding speech segments. This contrasts directly with the desire in unit selection to keep the natural unmodiſed speech units.

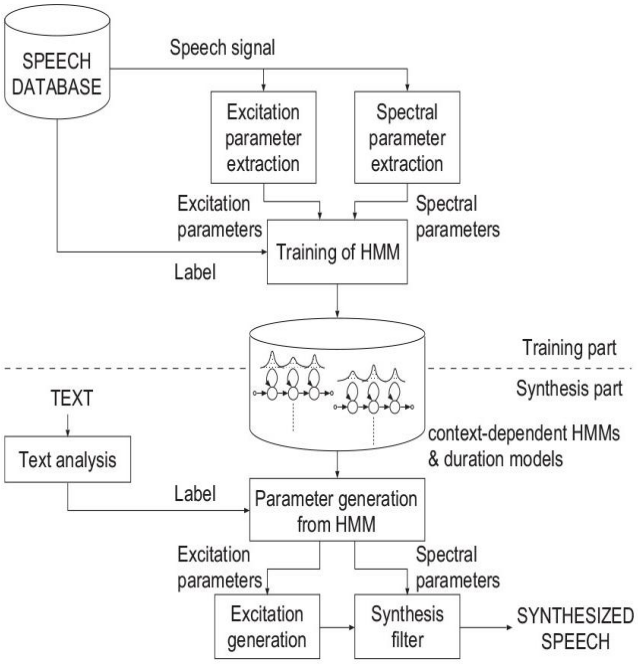

Figure 1

Secondly, state durations of the HMM are determined based on the state duration probability density functions. Thirdly, the speech parameter generation algorithm generates the sequence of mel-cepstral coefficients and log F0 values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated mel-cepstral coefficients and F0 values using the MLSA with binary pulse or noise excitation.

B. Advantages and disadvantages:

The biggest disadvantage of the HMM-based generation synthesis approach against the unit selection approach is the quality of synTraining of HMM context-dependent HMMs & duration models Training part Synthesis part Label Spectral parameters Excitation parameters Parameter generation from HMM TEXT Label Text analysis SYNTHESIZED SPEECH Excitation generation Synthesis filter Spectral parameters Excitation parameters Speech signal Spectral parameter extraction Excitation parameter extraction.

The HMM-based generation synthesis approach are

1) Its voice characteristics can be easily modiſed,
2) It can be applied to various languages with little modiſcation,
3) A variety of speaking styles or emotional speech can be synthesized using the small amount of speech data,
4) Techniques developed in ASR can be easily applied,
5) Its footprint is relatively small.

The voice characteristics in 1) can be changed by transforming HMM parameters appropriately because the system generates IV 1230 speech waveforms from the HMMs themselves.

III. SPEECH SYNTHESIS

The goal of this project is to provide a short but a comprehensive overview of Text-To-Speech synthesis by highlighting its digital signal processing component. First two rule-based synthesis techniques (formant synthesis and articulator synthesis) are explained then the concatenative synthesis is explored. Concatenative synthesis is simpler than rule-based synthesis, since there is no need to determine speech production rules. However, it introduces the challenges of prosodic modification to speech units and resolving discontinuities at unit boundaries. Prosodic modification results in artifacts in the speech that make the speech sound unnatural. Unit selection synthesis, which is a kind of concatenative synthesis, solves this problem by storing numerous instances for each unit with varying prosodies. The unit that best matches the target prosody is selected and concatenated. To resolve mismatches speech synthesis system combines the unit selection method with Harmonic plus Noise Model (HNM). This model represents speech signal as a sum of a harmonic and noise part. The decomposition of speech signal into these two parts enables more natural sounding modifications of the signal. Finally Hidden Markov model(HMM) synthesis combined with an HNM model is introduced in order to obtain a Text-To- Speech system that requires smaller development time and cost.

IV. SCOPE OF THE PAPER

This paper proposes a system that will create a revolution in a world of education by providing an easier way for visually impaired people to take tests just as normal students do. The system acts as a mediator who converts the responses that are given orally to the system to acceptable and needed format. When user gives response orally speech to text converter is invoked and it converts the response in text to mark the appropriate option of all. Similarly, other essential things like timer and result can also be heard. These are:

- Blind peoples are easily use web application itself.
- If there is no need help for other person.
- It provides voice output for all messages in web application.

Both the things can be invoked orally by just remembering few commands. There are different sections in the system. User can choose any one option through voice command. The whole system works on voice command so that everyone can use the same system using visually impaired people. In direct contrast to this selecting of actual instances of speech from a database, statistical parametric speech synthesis has also grown in popularity over the last few years. Statistical parametric synthesis might be most simply described as generating the average of some set of similarly sounding speech segments. This contrasts directly with the desire in unit selection to keep the natural unmodiſed speech units. A common speech database is provided to participants to build a

synthetic voice, the results from listening tests have shown that one of the instances of statistical parametric synthesis techniques called HMM-based generation synthesis (or even HMM-based synthesis) offers more preferred (through MOS tests) and more understandable (through WER scores) synthesis. Although even the proponents of statistical parametric synthesis feel that the best examples of unit selection are better than the best examples of statistical parametric synthesis, overall it appears that quality of statistical parametric synthesis has already reached a quality that can stand in its own right. The quality issue really comes down to the fact that given a parametric representation it is necessary to reconstruct the speech from those parameters.

V. SYSTEM ARCHITECTURE OVERVIEW:

A. Open New Account:

In this module the user open a new account for using the website. The most common and simple way of protecting a network resource is by assigning it a unique name and a corresponding password. Using this valid username and password, the user can view the account. A register user using this application, so the registration is must.

B. Training Module:

The user can hear and type the test using conversion process. There are two processes such as text to voice and voice to text. Using this technology the user can hear the Questioned. This process runs in between the User and System communication. These details are stored and retrieve from the database.

C. Text to Voice:

From the technology perspective, speech recognition has a long history with several waves of major innovations concatenating pieces of recorded speech that are stored in a database. The trained datasets are stored in the database just like as human dictionary. Most recently, the field has benefited from advances in deep learning and big data. Using this technology the receiver can receive the Question as sound format. So, the blind people can easily understand the Questions.

D. Performance Evaluation:

Communication is the main process in the world. Using internet facilities the user chat with other from anywhere. But blind people have a no chance to using these applications. This system introduces the email communication for blind people. It's a normal client and server communication through internet. It converts the voice to text and gathered the information.

E. Voice to Text:

In this module, Used to Choose the given answer option, the user can fix the answer to the each questions. Speech Synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. The trained datasets are stored in the database just like as human dictionary. Using these datasets, the data can be displayed.

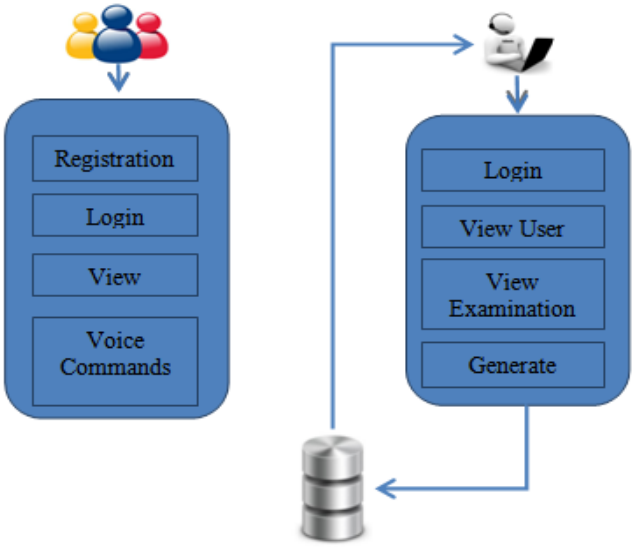

Figure 2

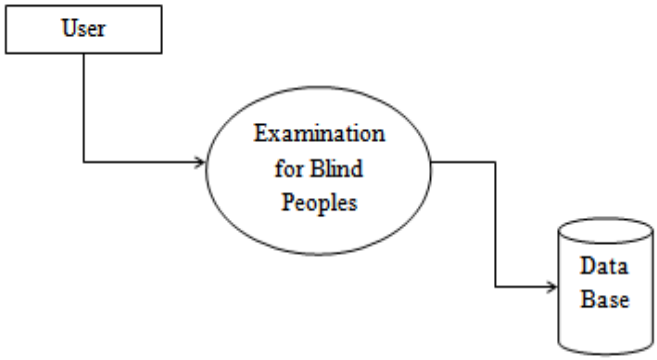

Figure 3

DFD Level 0 is also called a Context Diagram. It's a basic overview of the whole system or process being analyzed or modeled. It's designed to be an at-a-glance view, showing the system as a single high-level process, with its relationship to external entities. It should be easily understood by a wide audience, including stakeholders, business analysts, data analysts and developers.

VI. OVERVIEW OF WAVE FORM RECONSTRUCTION

The main aim of this paper is to propose a system that uses speech technology to provide students with access to information during exams. The key focus of application is to provide students with an ability to interact with the system through speech. application automates the examination process through reading out questions to the user and receiving their input orally. Application also provides accessories for other requirements, like knowing the time

remaining, during exams. Use of this application shall benefit students with:

- Learning disabilities
- Poor or limited motor skills
- Vision impairments
- Physical disabilities
- Limited English Language

The application will help the students with reading writing disabilities as well as sensory disabilities (blind or handicapped).

During synthesis, a vocoder based on minimum-phase or zerophase filter is often used together with the generated magnitude spectra to produce the synthesized output. Nevertheless, phase spectrum has been recently found to be essential for speech perception. The speech quality of vocoded outputs are found to be degraded from the original speech recordings. This may shed light on SPSS, where speech waveform with phase information in addition to the existing magnitude spectrum, is modeled.

In our work, speech signals are modeled by the corresponding magnitude and phase spectra, without the use of a vocoder. Consequently, reconstruction of speech waveform is facilitated. Commercial systems have exploited these technique to bring us a new level of synthetic speech. However, although certainly successful, there is always the issue of spurious errors. When a desired sentence happens to require phonetic and prosody contexts that are under represented in a database, the quality of the synthesizer can be severely degraded. Even though this may be a rare event, a single bad join in an utterance can ruin the listeners ƀow. It is not possible to guarantee that bad joins and/or inappropriate units do not occur, simply because of the vast number of possible combinations that could occur. However for particular applications it is often possible to almost always avoid them. Limited domain synthesizers, where the database is designed for the particular application, go a long way to making almost all the synthetic output near perfect. However in spite of the desire for perfect synthesis all the time, there are limitations in the unit selection technique.

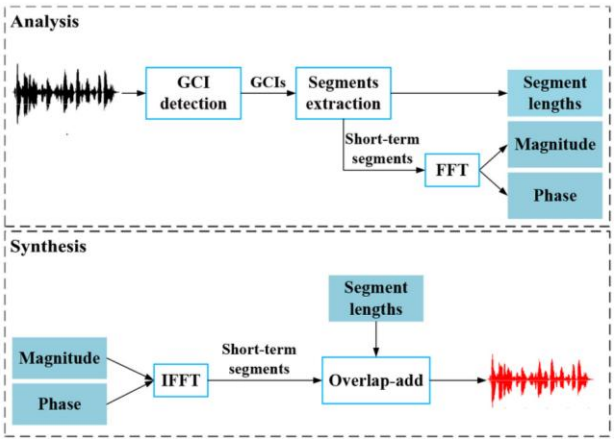

Figure 4

With a desire for more control over the speech variation, larger databases containing examples of different styles are required. But is limited by the amount of variations that can be recorded. In direct contrast to this selecting of actual instances of speech from a database, statistical parametric speech synthesis has also grown in popularity over the last few years. Statistical parametric synthesis might be most simply described as generating the average of some set of similarly sounding speech segments. This contrasts directly with the desire in unit selection to keep the natural unmodiſed speech units, but using parametric models offers other beneſts.

A. Experiment on Waveform Reconstruction:

Speech waveform in the test set of the corpus was analyzed and re-synthesized using our waveform representation framework and the three vocoders. The reconstructed speech waveform was then used for objective and subjective evaluations.

B. Objective Evaluation:

In the objective evaluation, we calculated the root mean square error between the reconstructed and original speech waveform signals in the voiced parts (RMSE voiced), the unvoiced parts (RMSE unvoiced) and the entire waveform), respectively. These voiced/unvoiced results from our framework and the three vocoders generally represent the performance on vowels/consonants respectively.

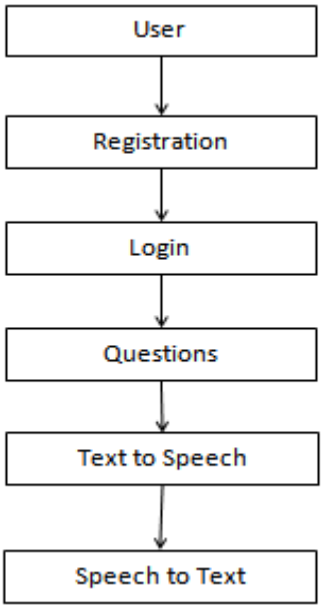

Figure 5

C. Subjective Evaluation:

Speech waveform are randomly selected from the reconstructed waveforms. Then a group of 20 subjects were asked to perform the preference test. We put the original waveform into X, while we put the waveform reconstructed using our framework and each of the three vocoders into A and B randomly. Each subject was asked to answer which one(A or B) is more similar to X. The third option Neutral means the subject has no preference on A or B. The ABX result is shown in Fig. 6. We can clearly see that the

reconstructed speech waveform using our framework is significantly preferred as compared with all of the three vocoders.

VII. EXPERIMENT ON WAVEFORM MODELING

In the baseline is used to vocode the speech waveform by a 25-ms moving window, and shifted every 5-ms. The generated magnitude spectrum from STRAIGHT was converted into LSP. The dimensionality of the input contextual label is 427.

The output feature contains Voiced/unvoiced flag a neural network with two BLSTM layers sitting on two feed forward layers with For our TTS system, features were extracted from the short term segments specified by GCI locations. The segment length is transformed into F0. The output feature comprises several components: voice/unvoiced and dynamic phase feature (257 dimensions), totally 300 dimensions.

The same network topology as baseline is used to train our TTS system. In the HMM-based generation synthesis approach, distributions for spectrum, F0, and duration are clustered independently. Accordingly, it has different decision trees for each of spectrum, F0, and duration. On the other hand, unit selection systems often use regression trees (or CART) for prosody prediction.

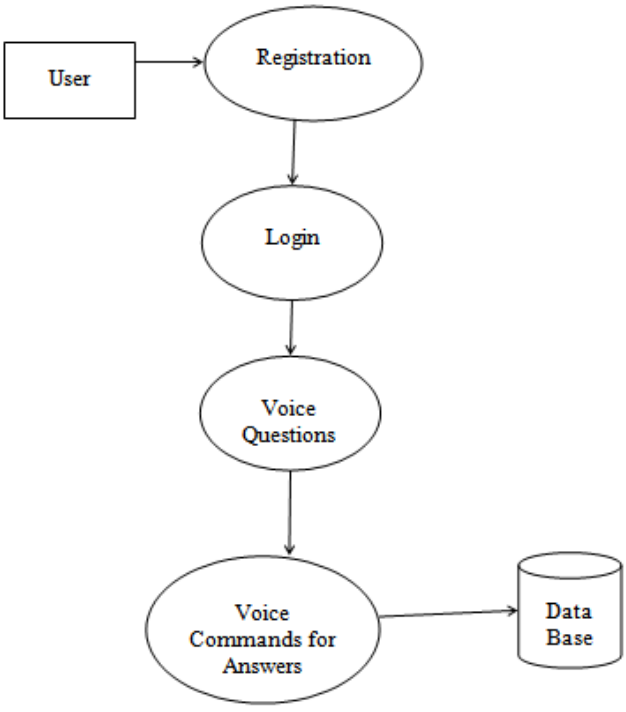

Figure 6

The decision trees for F0 and duration in the HMM-based generation synthesis approach are essentially equivalent to the regression trees in the unit selection systems. However, in the unit selection systems, leaves of one of trees must have speech waveforms: other trees are used to calculate target costs, to prune waveform candidates, or to give features for constructing the trees for speech waveforms. It is noted that in the HMM-based generation synthesis approach, likelihoods of static feature parameters and dynamic feature parameters corresponds to the target costs and concatenation costs, respectively. It is easy to understand, if we approximate each state output distribution by a discrete distribution or instances of frame samples in the cluster: when the dynamic feature is calculated as the difference between neighboring static features, the ML-based generation results in a frame-wise DP search like unit selection.

DFD Level 1 provides a more detailed breakout of pieces of the Context Level Diagram. It will highlight the main functions carried out by the system, as you break down the high-level process of the Context Diagram into its sub processes. We use a recently-emerging learning technique, well-suited for learning sequential events apart from long time lags of unknown size. Promising performance in various speech applications is observed. Our joint model of magnitude and phase is constructed. We employ line spectrum pair as the feature representation of magnitude spectrum. LSP, being an alternative LPC spectral representation, is robust and suitable for interpolation and modeling.

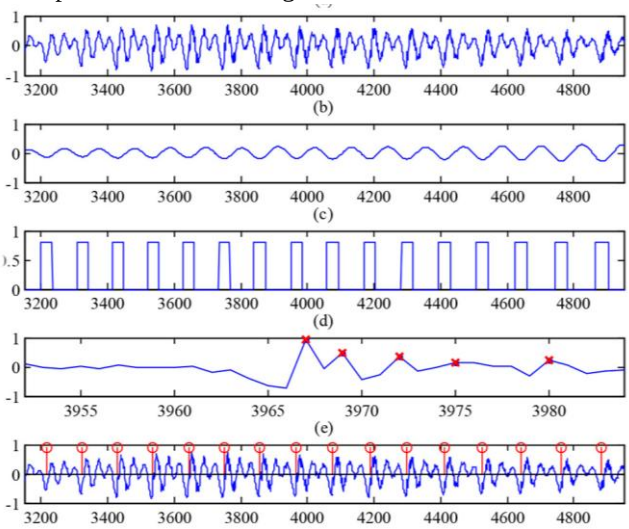

Figure 7

Statistical analysis of cepstral coefficients has shown that different emotional states are manifested in a speech signal in observed parameters of cepstral coefficients, histogram envelopes and together with other parameters, they may well be used for identification of individual emotions. The values given by numerical evaluation of obtained statistical parameters will be used for modification of the cepstral synthesizer digital approximation filter structure, including possible implementation in the Czech and Slovak TTS system based on cepstral description of speech inventory enabling expression of basic emotional speech styles. Results of the cepstral coefficient ranges and values statistical analysis are shown also in the form of histograms in a similar way as the spectral flatness ranges and values. This method can also be used for evaluation of emotional synthetic speech as a supplementary approach parallel to the listening tests.

VIII. WAVEFORM MODELING

State-of-the-art SPSS usually models the magnitude spectrum of speech signals and discards the phase spectrum.

Comparing the log magnitude spectrum with the dynamic phase spectrum, patterns of voiced and unvoiced portions are consistent and spectral patterns of individual speech sounds are quite similar in the log magnitude spectrum and the dynamic phase spectrum. This is important and useful for our joint modeling. On the contrary, there is no clear difference in the static phase spectrum for individual speech sounds. Statistical parametric synthesis might be most simply described as generating the average of some set of similarly sounding speech segments.

IX. RESULT

We have implemented proposed system with three individual section and features like timer and result for each one. For the implemented sections and subjects the system is running perfectly and flawlessly. models, when compared to standard unit selection, allow for general solutions, without necessarily requiring recording speech in all phonetic and prosodic contexts. The pure unit selection view requires very large databases to cover examples of all desired prosodic, phonetic and stylistic variation. In contrast statistical parametric synthesis allows for models to be combined and adapted thus not requiring instances of all possible combinations of contexts.There by our proposed application is suitable for use in real-time with high performance.

X. CONCLUSION

This paper would be a very useful one for every blind people and physically challenged to admire their talent easily through online exam like other humans. In our project we will be going to deliver an entire application for physically challenged people which can provide an interactive interface. Examinee can easily give exam by giving easy voice commands. Thus, physically challenged people can easily give exam like a common man without much difficulty. Through this they have been able to attend many exams in the future And also we will try to do as much as improvement in future as per the collection of feedback.

References

[1] Advanced .NET Re/moting 2nd Edition (Ingo Rammer and Mario Szpuszta, Apress, March 2005) ISBN: 1-59059-417-7
[2] Advanced .NET Remoting in VB.NET (Ingo Rammer, Apress, July 2002) ISBN: 978-1-59059-062-1
[3] ASP to ASP.NET Migration Handbook (Christian Nagel et al, Wrox, January 2003) ISBN: 978-1861008466
[4] Beginning Visual C# (Christian Nagel et al, Wrox, September 2001) ISBN: 1118314417
[5] Data-Centric .NET Programming (Christian Nagel et al, Wrox, December 2001) ISBN: 186100592X
[6] Professional .NET Network Programming 2nd Edition (Christian Nagel et al, Wrox, September 2004) ISBN: 1590593456
[7] Professional C# (Christian Nagel et al, Wrox, and March 2002) ISBN: 1430242337.
[8] Professional C# Web Services (Christian Nagel et al, Wrox, December 2001) ISBN: 1861004397
[9] Microsoft Visual C# .NET 2003 Developer's Cookbook (Mark Schmidt, Christian Nagel, etal, SAMS,October 2003) ISBN: 0672325802
[10] T. Nakatani, T. Irino, and P. Zolfaghari, "Dominance spectrum based v/uv classification and F0 estimation," in Proc. Eurospeech, 2003, pp. 2313–2316.
[11] K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in Proc. Eurospeech, 2003, pp. 2117–2120.
[12] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in Proc. Interspeech, 2014, pp. 1653–1657.
[13] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech communication, vol. 9, no. 5, pp. 453–467, 1990.
[14] T. Dutoit and H. Leich, "MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database," Speech Communication, vol. 13, no. 3, pp. 435–440, 1993.
[15] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," Speech and Audio Processing, IEEE Transactions on, vol. 9, no. 3, pp. 232– 239, 2001.
[16] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in Proc. ICASSP, 2015, pp. 4215– 4219.
[17] R. Maia, M. Akamine, and M. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in Proc. ICASSP. IEEE, 2012, pp. 4581– 4584.
[18] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," Speech and Audio Processing, IEEE Transactions on, vol. 3, no. 5, pp. 325–333, 1995.