Generating Expressive Degree of Emotion in Neutral Speech

Manjare Chandraprabha A, Shirbahadurkar Suresh D, Patil Prerna R

Abstract— This paper proposes a statistical phrase/accent model for speech synthesis. In recent years, the work on expressive speech has increased rather than basic emotions. Our aim is to obtain expressive speech from neutral speech. In this proposed method there are two components one is phrase and other is accent. Expectation-Maximization algorithm is used to train statistical speech data. The output generated by proposed method is compared with TD-PSOLA method. The results generated from proposed work is better than TD-PSOLA method.

Index Terms— Intonation Modeling, Accent/Phrase, Statistical parametric Speech Synthesis, TD-PSOLA.

I. INTRODUCTION

Most research has been worked on expressive speech synthesis which concentrates on basic emotions such as sadness, happiness, fear and anger. Along with basic emotions we also need other expressive speaking styles. In the area of expressive speech synthesis lot of efforts have been made recently. In this paper focus is kept on generating TTS system for novel reading, storytelling applications. Speech Synthesis is artificial production of human speech. A speech synthesizer is a computer based system used for artificial production of human speech with a completely synthetic voice output and can be in software or hardware. The obtained synthetic voice is a neutral voice. Naturalness and intelligibility are the most important qualities of speech synthesis system. Naturalness is the closeness of the obtained output sounds to the human speech. The overall aim of the speech synthesis is to generate natural sounding synthetic speech. Earlier attempts were made using rule- based synthesis to impart emotional effect to synthetic voice. To attract audiences the human storytellers modified their voices. They use to create variety of voice and sound effects themselves to engage listeners. TTS should give a listening experience that is equally attractive as human story tellers in an ideal case. Computer is used to narrate story in digital storytelling stories. Ideally digital storyteller should provide equal listening experience as of human storyteller. We want to create an application which will equal the voice of human storyteller by modifying the neutral voice parameters. To complete this target we need to concentrate on expressive style which is not available in the text to speech system these days. As we require more control on F0 for different speaking styles, emotion etc, we have to model F0 contour in order to synthesize the desired range of speech output.

Revised Version Manuscript Received on July 13, 2015.

Manjare Chandraprabha Anil, received M.E.(Electronics) degree from Marathwada University, Shivaji University, Aurangabad, Maharashtra, India.

Dr. Shirhahadurkar, Ph.D.(EC) degree from Marathwada University, Dr. B.A.M. University, Aurangabad, Maharashtra, India.

Prerna R. Patil, M.E.(Digital Systems) degree from Savitribai Phule Pune University, Maharashtra, India.

In this paper, we are going to compare two methods that is TD-PSOLA and Statistical Phrase/Accent Model for intonation modeling. We can see that the proposed work is better as compare to TD-PSOLA.

II. RELATED WORK

Researchers have done lot of work to improve speech synthesis. Welsey Mattheyses explains TD-PSOLA [1] technique for prosodic modification of speech signals, especially for pitch shifting and time scaling. The time domain pitch synchronized overlap-add algorithm (TD-PSOLA) is well known in the field of speech synthesis as it allows for high quality pitch and time scale modifications of stored speech segments and has a very low complexity and computational cost. However, it is also well known that the sound quality of TD-PSOLA modified speech is very sensitive to a proper positioning of the pitch marks that delimit the individual pitch epochs. Therefore, TD-PSOLA has been mostly used in applications such as text-to-speech synthesis, where the pitch marking can be done off-line and corrected manually. Later Andrej Ljolje, in his work used HMM for natural sounding pitch contour [2]. The mechanism consists of modeling a set of observations as a probabilistic function of a hidden Markov chain. It uses mixtures of Gaussian continuous probability density functions to represent the essential, perceptually relevant structure of intonation by observing movements of fundamental frequency in monosyllabic words of varying phonetic structure. High quality speech synthesis, using multipulse excitation, is used to demonstrate the power of the HMM in preserving the naturalness of the intonational meaning. In this the fundamental frequency contours are synthesized using a random number generator from the models. More recently, there was an approach towards fully automatic extraction of Fujisaki model parameters [3].It was for larger speech database of German. Fujisaki model was used for production process of F0. The focus was then diverted to Automatic intonation modeling [4] with INTSINT. MOMEL algorithm was used for automatic derivation of target points. Then there was voice conversion method [5] based on analysis and transformation of the characteristics that define a speaker's voice. **A** PSOLA based method is used for transformation of pitch, intonation patterns and speaking rate. In this de-convolution of vocal tract is used for modeling and mapping of glottal pulse. Modeling of intonation variability [6] was done with HMM which also used CART. The focus was then moved towards CART method, where modeling of intensities of syllables was done using CART [7]. To further improve prosody accent group modeling [8] for speech synthesis was introduced. Intonation modeling was based on SPSS (Statistical Parametric Speech Synthesis). Resynthesis error

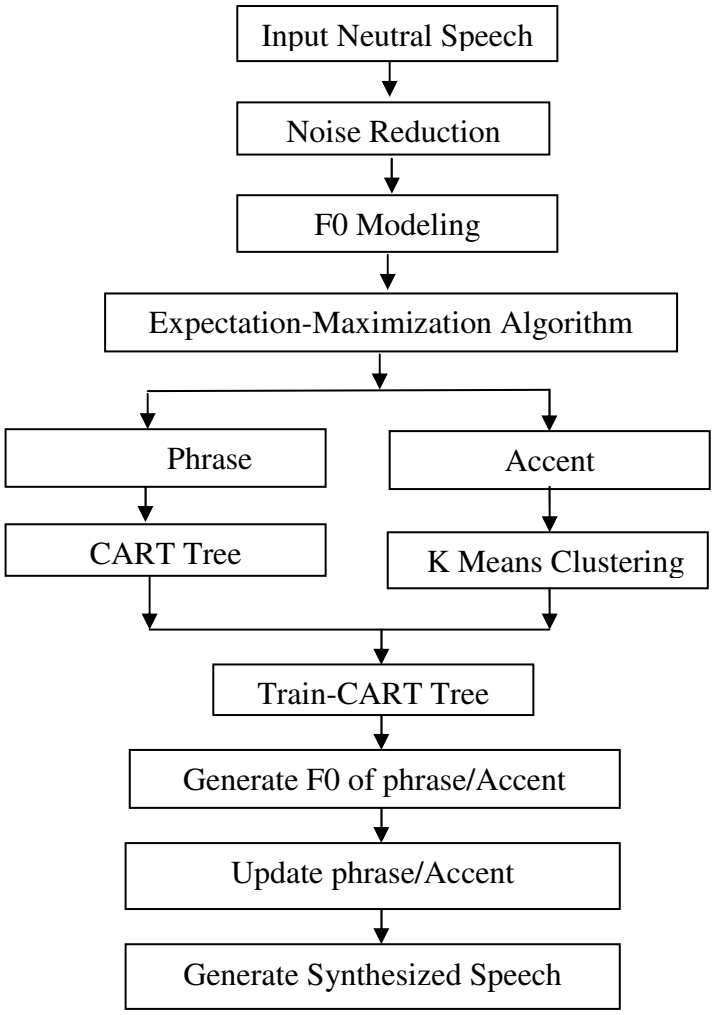
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

minimization algorithm is used for automatic accent group extraction .

III. BLOCK DIAGRAM OF PROPOSED SYSTEM

A. Architecture of Proposed System

In this project our main goal is to present efficient framework for emotion generation in neutral speech by using accent/phrase model. Below figure is showing the architecture of proposed work.



B. Block Diagram Description

Firstly both the input speech that is neutral speech and emotion speech is given to the system. It will read both the audios and plot the graphs. Then the preprocessing of speech signal takes place where the noise is removed from the speech to make it smooth and improve the efficiency while generating the emotions in neutral speech. Then after obtaining processed speech it is passed further for F0 modeling. In F0 modeling, conventional F0 estimation of input neural signal takes place. This can be further used in constrained component extraction algorithm. In this constrained component extraction algorithm is the Expectation-Maximization algorithm which is used to statistically train the components from speech data. In statistics, an iterative method of finding maximum likelihood or maximum estimates of parameters is called as expectation–maximization (EM) algorithm. It is an iteration between (E) and (M) step, in which (E) step creates function for expectation of log-likelihood and (M) step computes the parameters which maximize the expected log-likelihood which is there in (E) step. In the next E step, the estimated parameters are then used to determine the distribution of the latent variables. This algorithm is used to train the phrase /accent components. An initial estimate of phrase command is used. As an approximation of the phrase component we will use minimum value of F0 over a syllable. The residual is consider as accent for each syllable. At this stage, for phrase component CART tree is build and for accent components k-means clustering is used. K-means clustering is a method of vector quantization, used mostly in signal processing. k-means clustering partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The expectation-maximization mechanism allows clusters to have different shapes whereas , it uses cluster centers to model the data. k-means clustering tends to find clusters of comparable spatial extent. The system is robust to utter specific artifacts of the speaker or pitch detection routines as the components are trained over the entire training data. It gives the model, more degrees of freedom as the constraints are chosen to be minimally assuming and are generic across languages, speakers or speaking styles, After the intermediate models are built (phrase CART tree and accent codebook), a new estimate log(F0) is reconstructed. Then the mean reconstruction error over each syllable is added to the previous baseline and residuals are recomputed. Till it obtains the minimum F0 reconstruction error this procedure is repeated. The parameters that give the best reconstruction error are chosen as the optimal phrase and accent components. After this we update both phrase and accent values to generate synthesized speech.

C. The main steps and there explanation

Step 1: Extract F0 Features using F0 Modeling Specification: F0 modeling is does the task of extracting the frequency from the input speech in which pitch is recognized and leaving salient part. This function of F0 modeling helps to focus on synthesizing the only voice part from the input speech signal. Therefore in most of cases F0 modeling is referred to use as speech extraction function.

Step 2: For all F0 features do extract phrase component and accent component by using two below equations:

Phrase= min {F0}

Accent=tilt (F0 – phrase)

Specification: These are two main components on which our algorithm further working for generating synthesized speech.

Step 3: Generating accent codebook and phrase codebook while error is more. Below are functionalities used: This phase is also known as constrained component extraction algorithm

1) A k-means clustering is performed to identify the representative shapes of accents over speech signal. This is done by performing the task iteratively. This is called training of accent codebook using k-means clustering still there is more error. Output of k-means is referred as codebook. This trained output is known as accent codebook.

2) After that the CART tree is applied on two kinds of features

- low level features extracted by k-means clustering in codebook.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

- high level/long range features extracted by F0 modeling in phrase

The CART tree functionality is used to train High-pass filter speech in order to remove possible low frequency fluctuations and generate the new phrase codebook. The output of CART tree is known as phrase codebook.

Specification: The main aim of this phase is to get less error between new F0 and existing F0.

Step 4: Iterative optimization algorithm

In this stage after generating accent codebook and phrase codebook, we then apply iterative optimization algorithm to generate new F0 model. The aim of this phase is to repeat the below process till an objective criterions is met which is the minimum F0 reconstruction error. The parameters that give the best reconstruction error are chosen as the optimal phrase and accent components.

IV. TECHNIQUES OF SPEECH MODIFICATION

A. CART Tree Method

Classification and regression trees are used for constructing prediction models from data. CART-based synthesis is based on CART-based model and is based on machine learning method. CART's have been used in prediction of prosody, such as, prosody phrase boundaries, duration etc. The CART is a simple binary decision tree made by feeding the attributes of the feature vectors from top node, and it passes through the arcs which represents the constraints. By passing the feature vector through the tree the intensity of a segment is predicted so as to minimize the variance at each terminal node. It can capture a good amount of statistical variation into the model with small memory consumption. The algorithm usually guarantees that the tree fits the training data well. The performance of the CART model depends on the coverage of the training data.

B. K-means clustering

k-means clustering is a method of vector quantization, mostly used in signal processing and data mining. k-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

C. TD-PSOLA

The time domain pitch synchronized overlap-add algorithm (TD-PSOLA)[1] is well known in the field of speech synthesis. TD-PSOLA allows high quality pitch and time scale modifications of stored speech segments and has a very low computational cost and complexity. It is also well known that the sound quality of TD-PSOLA modified speech is very sensitive to a proper positioning of the pitch marks that delimit the individual pitch epochs. Therefore, TD-PSOLA has been mostly used in applications such as text-to-speech synthesis, where the pitch marking can be done off-line and its correction can be done manually. Time-domain TD-PSOLA is the most popular PSOLA technique. However, TD-PSOLA does not allow any other form of modification (e.g., spectral). It can only be used for time- and pitch-scaling.

V. RESULTS

Here figure 1 and 2 showing the speech signals for original neural and its emotional sentence. We required emotional speech signal of same sentence for comparative study purpose. We are going to compare our proposed work with TD-PSOLA method. All results are listed out below one by one.

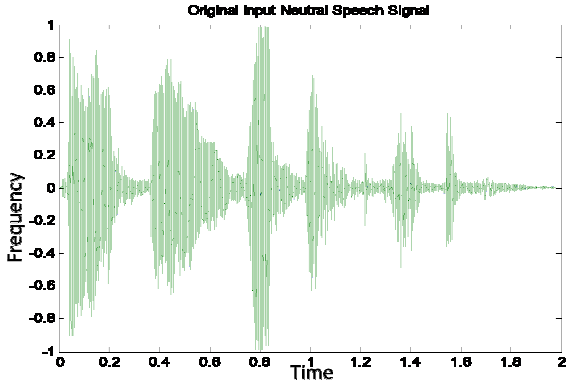


Figure 1: Input Neutral Speech Signal for "I am happy"

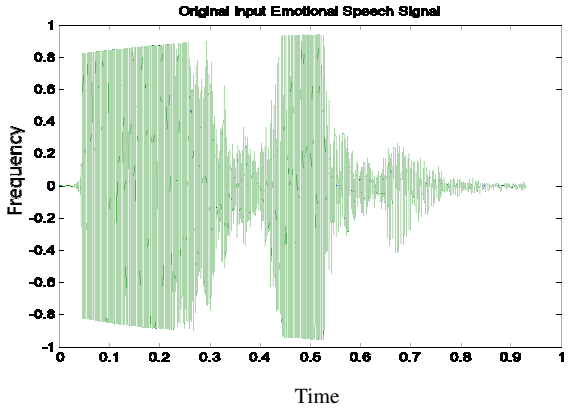


Figure 2: Input Emotional speech signal for "I am happy"

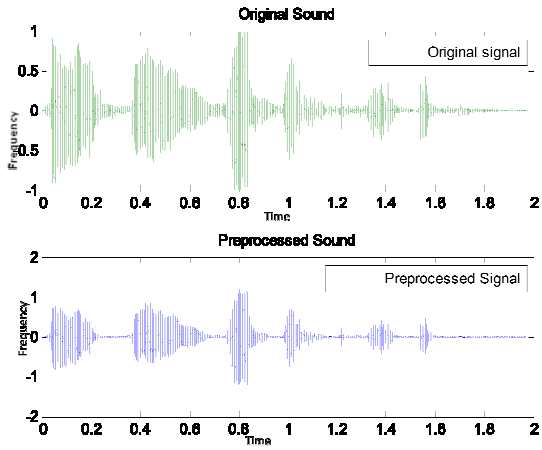


Figure 3: Preprocessed signal

Figure 3 is showing the preprocessed single in which, signal becomes smooth, noises are reduced for further processing, and this improves the efficiency while generating the emotions in neutral speech. Figure 4 below is showing conventional F0 estimation of input neural signal. This can be further used in constrained component extraction algorithm

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

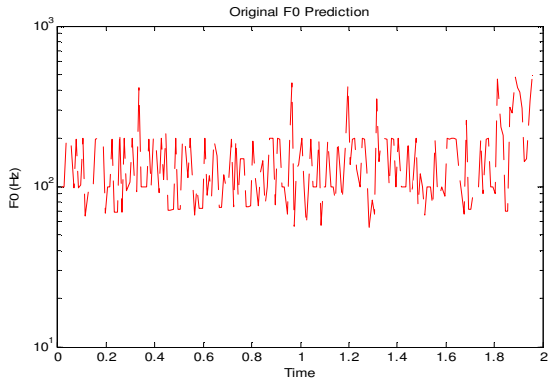


Figure 4: Original F0 Prediction of input signa

Below figure 5 and 6 showing the comparative results for proposed work.

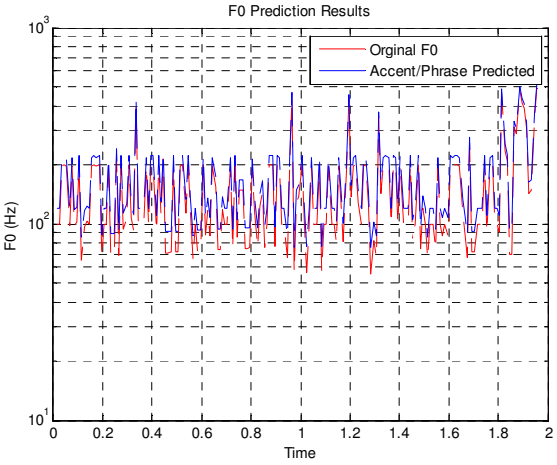


Figure 5: Comparative analysis of speech estimation

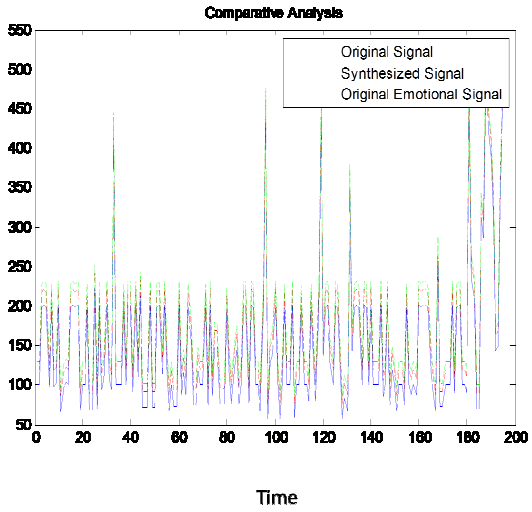


Figure 6: Comparative analysis of speech synthesis results against original neutral and emotional signal

VI.DISCUSSIONS

In this section we will present comparative analysis of TD-PSOLA and proposed method in terms of two performance metrics such as Correlation (CORR) and Root Mean Square Error (RMSE) with below significance. Correlation: This metrics represents the similarity between source voice and modified voice using cross correlation. Therefore, for efficient method it is required that less the correlation more the better method. The results showing the same for TD-PSOLA is more and for proposed method is less. RMSE: Root Mean Square Error is exactly opposite to Correlation metrics. This represents the difference between original voice and modified voice. Therefore for efficient method it is required that RMSE should be more as compared to TD-PSOLA method. We have represented 10 input neutral speech signals and based on observations below table 1 showing results of RMSE and CORR for TD-PSOLA method and proposed accent/phrase method

Table 1: Comparative Study

File_Name	TD-PSOLA		Accent/Phrase	
	CORR	RMSE	CORR	RMSE
a.wav	0.5122	10.4257	0.3387	14.0029
b.wav	0.6514	10.1015	0.4545	14.0055
c.wav	0.7146	10.0967	0.4836	14.0033
d.wav	0.7890	10.0599	0.4415	14.0003
e.wav	0.7315	10.0907	0.2671	14.0049
f.wav	0.7093	10.0674	0.5449	14.0006
g.wav	0.6210	10.0928	0.4190	14.0046
h.wav	0.8386	10.1162	0.4252	14.0015
i.wav	0.7613	10.0831	0.6273	14.0016
j.wav	0.8381	10.0930	0.5341	14.0020

As per the significance stated above, our results are better and improved as compared to TD-PSOLA method.

VII. CONCLUSION AND FUTURE WORK

As per the aim of this project, results obtained for speech synthesis are better and improved as compared to TD-PSOLA. There is new framework for speech synthesis called as F0 estimation using accent/phrase model. This can be also used for text-to-speech system. As required the results obtained for Correlation in case of proposed work is less as compare to TD-PSOLA and for Root Mean Square Error it is more. For future work we will suggest to work on speaker verification with speech synthesis under real time environment.

REFERENCES

[1] Welsey Mattheyses ,Werner Verhelst and Piet Verhoeve, "Robust pitch marking for prosodic modification of speech using TD-PSOLA".
[2] Andrej Ljolje ,Frank Fallside,"Synthesis of natural sounding pitch contours in isolated utterances using HMM", 1986.
[3] Hansjorg Mixdorff, "A novel approach to the full automatic extraction of Fujisaki model parameters", 2000.
[4] J.A.Louw and E.Barnard," Automatic intonation modelling with INTSINT", 2001.
[5] Dimitrios Rentzos , Saeed Vaseghi , Emir Turajlic , Qin Yan, Ching-Hsiang, "Transformation of speaker characteristics for voice conversion", 2003.
[6] Cedric Boidin,Olivier Boeffard, "Modeling Intonation Variability with HMM for Speech Synthesis", 2004 .
[7] Jing Zhu, Yibiao Yu, "Intonation and prosody conversion for expressive Mandarin speech synthesis", 2012.
[8] Gopala Krishna Anumanchipalliyz Lu´ıs C. Oliveiraz Alan W Blacky,"Heuft Accent group modeling for improved prosody in statistical parametric speech synthesis", 2013.
[9] Jinfu Ni, Shinsuke Sakai, Tohru Shimizu, and Satoshi Nakamura," Prosody modeling from tone to intonation in Chinese using the fundamental F0 model", 2008.
[10] Jinfu Ni, Yoshinori Shiga, and Chiori Hori,"Superpositional HMM-based intonation synthesis using a fundamental F0 model", 2014.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Manjare Chandraprabha Anil received B.E.(Electronics), M.E.(Electronics) degree from Marathwada University, Shivaji University, in 1992, 2005, respectively. Presently She is working on Marathi TTS system development. She authored/coauthored 10 publications and 6 International Conferences.

Dr. Shirhahadurkar received B.E.(EC), M.E.(EC), Ph.D.(EC) degree from Marathwada University, Dr. B.A.M. University, Aurangabad, DOEACC Aurangabad, in 1991,1998, 2010 respectively. He authored 12 International Journals, 14 International Conferences, 3 books. His area of Interest is Speech processing.

Prerna R. Patil received B.E.(Electronics and Tele Communication), M.E.(Digital Systems) degree from Savitribai Phule Pune University, in 2010, 2015, respectively. She authored 1 publication in International Conference.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.