IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.
PP 61-66
www.iosrjournals.org

Emotional speech synthesis by manipulating Speech parameters

Prashul B. Kamble[1], Shaila D. Apte[2]
[1](E&TC Department, Rajarshi Shahu College of Engineering, Tathwade, Pune,India)
[2]((E&TC Department, Rajarshi Shahu College of Engineering, Tathwade, Pune,India

ABSTRACT: *Emotion recognition from human speech signal has gained increasing attention by researchers in past few years. It is the process in which one of the emotional states of human is assigned to well known categories of emotion like happiness, sadness, anger, etc. Along with analysis phase, synthesis or conversion of emotion from one state to another state is also the attractive area for researchers. The present paper deals with the same set of analysis-synthesis approach. We have focused our work on two emotions namely happy and anger. We have concentrated on pitch and the vocal tract parameters in time domain namely LPC. The present work tries to track and modify the pitch marks and LPCs of each segment for emotion conversion. An all pole model system is designed in order to synthesize emotional speech. The modified filter is excited by modified source system. Frame by frame synthesis of speech improves the quality (Naturalness and Intelligibility) of speech signal. The correlation between the original and synthesized emotional speech signal is computed for happy and anger speech signal and it is 87.28% and 85.82% respectively. The percentage perception result of synthesized emotional speech are 73% and 85% for happy and anger speech signal respectively.*

Keywords *- Analysis and synthesis, Emotional speech synthesis, LPC, Pitch, LP residual, performance measure.*

I. INTRODUCTION

Speech synthesis is a process or technique in which human speech signal is artificially generated by a computer system. This computer system may be implemented in hardware or software. The two important qualities i.e. naturalness and intelligibility of synthesized speech signal are judged by comparing it with human speech. But as there is a lot of difference in artificial and natural speech production, researchers are on their way to minimize this difference and hence this fact motivates people to the area of speech synthesis for more than a decade. But still many aspects of emotional speech synthesis are largely unexplored and it is a challenging task for researchers.

After the extensive literature review, it is found that both psychology and speech tells the information regarding emotion in speech. This information may be a combination of prosodic, tonal and spectral information. Along with these speaking rate and stress distribution are found to play important role. Jiří Přibil et al. (2012) focused their attention on formant features in order to describe vocal tract characteristic in emotional and non-emotional speech. They have experimented using Czech and Slovak speech material which was extracted from the stories performed by actors. Elif Bozkurt et al. (2010) investigated behavior of different emotional speech on line spectral frequency (LSF). They made a conclusion that in terms of spectral features, LSF's are more emotion specific over MFCC. Agarwal et al. (2010) focused their attention on only emotion specific word. Their emotion conversion model was based on segmentation of the spoken utterance into words. The segmentation of word was done by using word boundary detection technique. Degaokar et al. (2011) examines the behavior of emotions speech at different frequency sub band using WPT. They also performed emotion modeling and emotion conversion using transition in sub bands of WPT. Koolagudi et al. (2010) have introduced how epoch parameters changes for different emotions. Various epoch parameters like strength of epoch, instantaneous frequency, sharpness of epoch, slope of strength of epoch are used as features for classification of emotions. Nicholos et al. (2010) used prosodic as well as phonetic parameters like power intonation pattern, etc in order to information to classify emotions. The classification is done using neural network classifier. Haojie Zhang et al. (2012) have performed experiment on the combination of pitch and formant. They first analyze the prosodic features and synthesize the emotional speech. TD-PSOLA method was used to modify prosodic parameters.

At the time of speech production we utter different words. Different words are uttered because of change into resonant modes of the vocal cavity along with stretching of vocal cords for modifying the pitch period for different vowels. For emotion, pitch change plays an important role. When the emotions are pleasant, the larynx and pharynx get expanded and vocal tract wall gets relaxed, whereas in case of unpleasant emotion, the larynx and pharynx get constricted and vocal tract wall becomes tensed. These facts motivate us to

deal with vocal tract and vocal cord parameter for experimentation. Hence after doing experimentation on these parameters (LPC modification and Pitch modification), we have successfully synthesized emotional speech.

The paper is organized as follows. Section 2 briefs the databases used. System implementation is explained in section 3. Analysis of parameters is discussed in section 4. Synthesis of emotional speech is discussed out in section 5. Paper is concluded with directions for future work

II. DETAILS OF DATABASES

In order to obtain the good quality of synthesized speech, the original speech signal must be noise free. For this we have used Berlin emotional speech database (EMO-DB). The database is freely accessible to researchers. EMO-DB contains the emotions from MPEG-4 standard. The sentences are recorded with anger, joy, disgust, surprise, sadness, neutral and fear. Professional actors were used for every utterance with five female actors. The data base is available in 16 bit format and 16 KHz sampling frequency with mono recording.

III. IMPLEMENTATION OF SYSTEM

The synthesis of emotional speech is achieved by analysis-synthesis method. The time domain vocal tract parameters like LPC are computed for neutral, anger and happy emotion. Along with this the prediction error for every emotion is computed and excitation information is extracted from it. A brief discussion about the introduced terms is as follows.

Linear Predicting Coefficient (LPC): The production of speech signal is mathematically described as convolution between the excitation source and the vocal tract parameters. In order to study these components, they must be separated from the speech signal. For deconvolution of these components, method like homomorphic analysis is used. One of the homomorphic analysis methods that is cepstral analysis can be used. But the cepstral analysis deconvolves these two components in frequency domain. So it becomes computationally complex process. Hence these source and system components are extracted from time domain itself to reduce the computational complexity. Hence linear prediction methods are developed to achieve this task.

The basic concept behind linear prediction of speech signal is that the prediction of current samples of speech signal from past available speech sample. The past available samples m decides the order of predictor. The current predicted sample can be represented as linear combination of past m samples as

$$\hat{s}(n)=-\sum_{k=1}^{m} [\![a_k\, s(n-k)]\!]$$

Here, a_k's are the LPC and s(n) is windowed speech signal

$$s(n)=x(n).w(n)$$

Generally, the vocal tract system is considered as an all pole filter in which pole locations are formants. If the factors at denomination are multiplied, we get transfer function as

$$H(z)=z^{m}/(z^{m}+ [\![a_1 z]\!] ^{(m-1)}+a_2 z^{(m-2)}+\cdots+a_{(m-1)}\, z+a_m)$$

Where a_1, a_2…etc are the LPC Coefficients.

Various methods for computation of LPC are as follows:

- Autocorrelation method.(Levinson–Durbin algorithm)
- Covariance method.
- Lattice structure method. (Burg's algorithm)

COMPUTATION OF LP RESIDUAL FROM LP ANALYSIS:LP residual is computed as the difference between the predicted speech signal and original speech signal. This is referred as LP residual signal. Hence the prediction error is given as

$$e(n)=s(n)-\hat{s}(n)$$ from (1) , we have

$$e(n)=s(n)+\sum_{k=1}^{m} [\![a_k\, s(n-k)]\!]$$

ESTIMATION OF PITCH FROM LP RESIDUAL: When the predicting coefficients are chosen in order to minimize the prediction error the generated error signal approximately represents the excitation signal. The residual energy consists of the component of signal which is not predicted from past samples. This kind of situation occurs for periodic excitation. If the residual signal is observed carefully, we can see the peaks present at pitch period interval representing excitation information. This excitation information is extracted from the error signal and it is used as excitation signal. To obtain this tract the peaks in residual signal and pulses so obtained are used as excitation to generate speech signal. The experimentation is done as follows. Firstly the entire speech signal is divided into number of frames each of size equal to pitch interval. The LPC for emotion specific word are extracted for each three class of emotion. Along with this excitation information is also

extracted from LP residual. After comparing with reference and target emotional speech, a modification factor is evaluated. This modification factor is then used for synthesis purpose.

IV. ANALYSIS PHASE

LPC and LP residual of speech signal with different emotions (e.g. Neutral, Anger and Happy) are computed in analysis phase. The computed LPC and LP residual are then compared with target emotional speech to find modification factor.

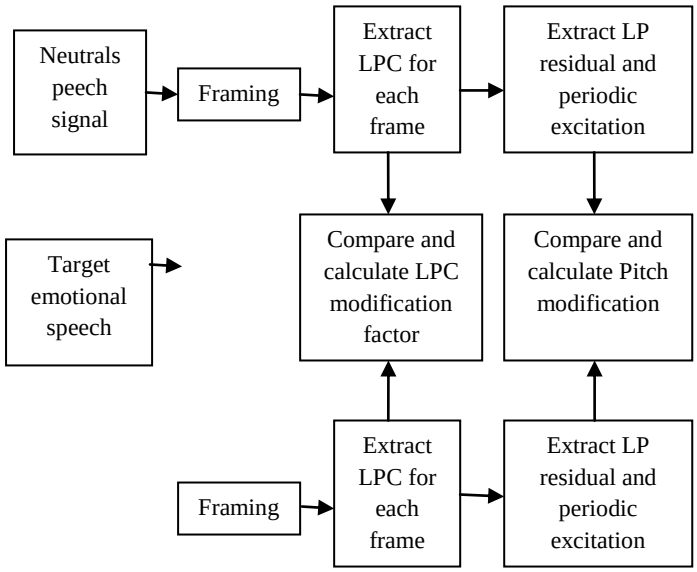

Figure 1: Analysis of speech parameters

The neutral speech signal is taken as reference speech signal. The entire speech signal is divided in number of frames for both neutral and target emotional speech. LPC coefficients for each frame is computed and compared with target emotional speech. On comparison, modification factor is computed for synthesis phase. The extracted LPC for a single frame of three different emotions is shown in figure 2.

12 LPC coefficients for three different emotions are shown in figure 2. It is found that the coefficient number 2, 3 and 4 playing important role in expressing the speech as happy. Hence comparing these three coefficients' with respective neutral coefficient, modification factor is computed. On similar lines LPC coefficient for Neutral and Anger speech signal are compared. It is found that the coefficient number 2 and 3 are playing important role in expressing speech as angry. Hence comparing these two coefficients with respective neutral coefficients modification factor is computed.

Now, to extract LP residual the predicted speech signal is subtracted from the original speech signal. Figure 3 shows residual signal for a speech segment.

The LP residual signal represents an error signal which is obtained by LP analysis. If we observe figure 3, we can see that the signal is noisy in nature. The corresponding pitch marks are characterized by sharp and periodic discontinuity. This type of discontinuity causes a large amount of error in the residual signal. The periodicity in error signal gives the pitch information for the segment of speech. The extracted pitch information is shown in figure 4.

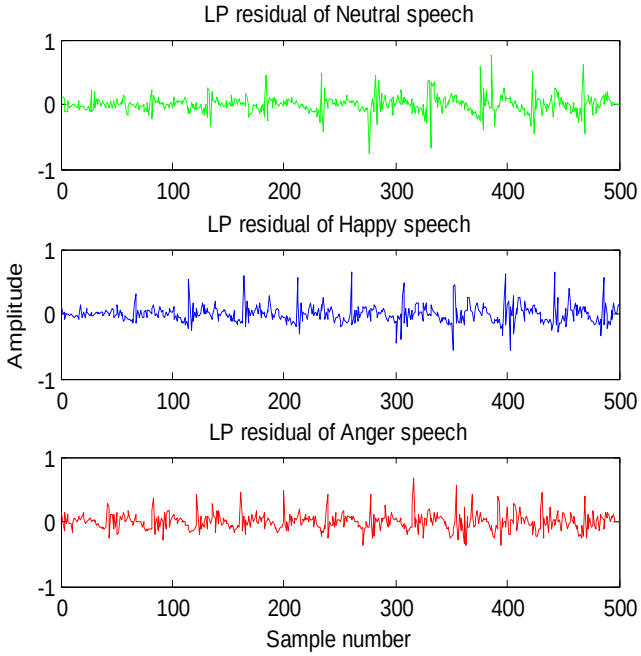

Figure 2: LP residual for three different emotional speech

Computation of modification factor: In the analysis phase, it is found that the second and third LPC plays an important role in anger expression where as second, third and fourth coefficient plays important role in happy emotion. The procedure for finding LPC modification factor is described as follows.

- The LPC parameter for a frame of neutral speech and target emotional speech is calculated.
- The LPC modification factor is calculated as ratio between two quantities as follows.

LPC modification factor=LPC for target speech/LPC for neutral speech

This modification factor is then applied to the neutral speech to add emotions to it. On the similar lines the modification factor for pitch information is calculated as

Pitch modification factor =Mean pitch period for target speech/Mean pitch period for neutral speech

The computed modification factors are as follows:

Parameter	Anger	Happy
Pitch period	0.71	0.89
Pitch amplitude	2.54	1.25

V. SYNTHESIS PHASE

Speech synthesis is a process of computer generation of speech. There are various methods for this purpose. The method which is adapted here is LP synthesizer.

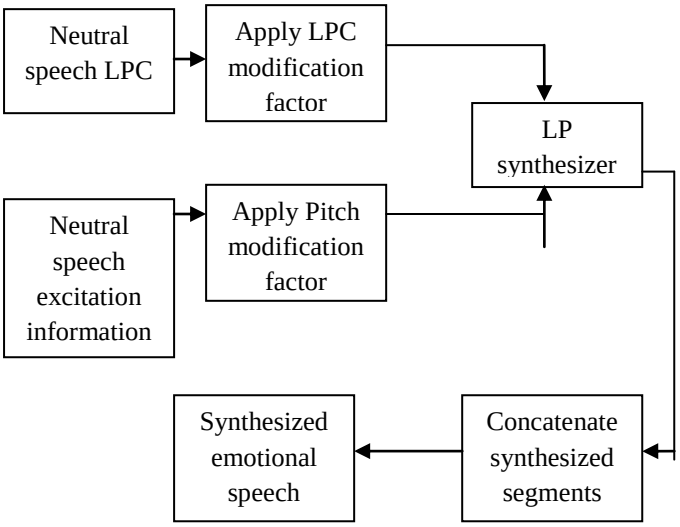

Figure 3: Block schematic for emotional speech synthesis using LP synthesizer

Synthesis of anger and happy speech signal:
For synthesis of anger speech signal, we have the neutral value of LPC coefficient and neutral excitation information. Along with this we have modification factor for both the parameters. The second and third neutral LPC coefficients are modified by modification factor respectively and an all pole transfer function representing a vocal tract filter is modeled. The excitation signal of neutral speech signal is also modified both in amplitude and duration respectively. The original and synthesized frame of anger speech is shown in figure 6.

On similar lines, modification in second, third and fourth LPC coefficient of neutral speech signal is done by a factor of respectively and excitation signal of neutral speech signal is also modified both in amplitude and duration by 1.25 and 0.89 respectively. The modification of these parameters results into happy speech signal. The original and synthesized frame of happy speech signal is shown in figure 7.

VI. Results And Discussion

The complete sets of utterances were presented in front of expert listeners in order to carry out subjective listening test. They were not only asked to identify the emotion but also to rate synthesized speech on the basis of naturalness and intelligibility. The confusion matrix for identification of emotional speech is summarized in table 2.

	Neutral	Anger	Happy
Neutral	90/-	0/-	10/-
Anger	0/-	87/85	13/-
Happy	16/11	0/-	84/73

Table 2: Confusion matrix for EMO-DB database

The test score for synthesized emotional utterance are 85% for angry and 73% for happy. These scores are given on the basis of quality of synthesized speech

VII. CONCLUSION AND FUTURE WORK

Literature regarding various emotion analysis and synthesis techniques is done and the approach is taken to use LPC as a new parameter for the proposed method. Various experimentations are carried out on standard Berlin database (EMO-DB) and our own created database. The conclusions are as follows:
Pitch modification of neutral speech results into emotional speech. Along with pitch modification, LPC coefficients are modified by the estimated modification factor respectively, happy emotional speech signal comes as output, whereas modification of LPC coefficients respectively results in angry emotional speech. Recognition rate of anger speech is (85%) higher than happy speech (73%). Pitch of anger emotional speech is

higher than neutral and happy emotional speech. Whereas, Pitch of happy emotional speech is higher than neutral emotional speech. i.e. $\text{Pitch}_{\text{Anger}} > \text{Pitch}_{\text{Happy}} > \text{Pitch}_{\text{Neutral}}$

ACKNOWLEDGEMENTS

It is my humble pleasure to get this opportunity to thank to my beloved and respected guide Prof. Dr. Mrs. S. D. Apte, who imparted valuable basic knowledge of Electronics specifically related to Speech Processing during the complete execution of this work.

REFERENCES

[1]. M. Schroeder, "Emotional speech synthesis: A review," *Proceedings of Eurospeech 2001*, 1, pp. 561–564, Aalborg, Denmark, 2001.
[2]. Dimitrios Ververidis, Constantine Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication, Volume 48, Issue 9, September 2006, Pages 1162-1181, ISSN 0167-6393, http://dx.doi.org/10.1016/j.specom.2006.04.003.
[3]. Bozkurt, Elif, Engin Erzin, and Çigdem Eroğlu Erdem. "Use of line spectral frequencies for emotion recognition from speech." *20th International Conference on Pattern Recognition (ICPR), 2010*. IEEE, (pp. 3708-3711). 2010.
[4]. Koolagudi, S.G.; Reddy, R.; Rao, K.S., "Emotion recognition from speech signal using epoch parameters," *International Conference on Signal Processing and Communications (SPCOM),2010*, IEEE, vol., no., pp.1,5, 18-21 July 2010
[5]. Ziolko, M.; Jaciow, P.; Igras, M., "Combination of Fourier and wavelet transformations for detection of speech emotions," *7th International Conference on Human System Interactions (HSI), 2014*, IEEE,vol., no., pp.49,54, 16-18 June 2014
[6]. Han Zhiyan; Wang Jian, "Speech emotion recognition based on wavelet transform and improved HMM," *25th Chinese Control and Decision Conference (CCDC), 2013*, IEEE,vol., no., pp.3156,3159, 25-27 May 2013
[7]. Jianhua Tao, Member IEEE, Yongguo Kang, and Aijun Li, "Prosody conversion from neutral speech to emotional speech", *IEEE transaction of audio, speech, and language processing*, vol. 14, no. 4, July 2006
[8]. Qin, Long, Zhen-Hua Ling, Yi-Jian Wu, Bu-Fan Zhang, and Ren-Hua Wang. "HMM-based emotional speech synthesis using average emotion model." In *Chinese Spoken Language Processing*, pp. 233-240. Springer Berlin Heidelberg, 2006.
[9]. Inanoglu, Zeynep, and Steve Young. "Emotion conversion using F0 segment selection." In *INTERSPEECH*, pp. 2122-2125. 2008.
[10]. Khorinphan, C.; Phansamdaeng, S.; Saiyod, S., "Thai speech synthesis with emotional tone: Based on Formant synthesis for Home Robot," *Third ICT International Student Project Conference (ICT-ISPC), 2014*, IEEE, vol., no., pp.111,114, 26-27 March 2014
[11]. R. Muralishankar, A.G. Ramakrishnan, P. Prathibha, Modification of pitch using DCT in the source domain, Speech Communication, Volume 42, Issue 2, February 2004, Pages 143-154, ISSN 0167-6393, http://dx.doi.org/10.1016/j.specom.2003.05.001.
[12]. V. Degaonkar and S. Apte, "Emotion modeling from speech signal ased on wavelet packet transform," *International Journal of Speech Technology*, Springer, vol. 16, no. 1, pp. 1–5, 2013.
[13]. Inanoglu, Zeynep, and Steve Young. "A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality." In *INTERSPEECH*, pp. 490-493. 2007.
[14]. David Suendermann, Harald Höge, and Alan Black. "Challenges in speech synthesis." In *Speech Technology*, pp. 19-32. Springer US, 2010.
[15]. Rank, Erhard, and Hannes Pirker. "Generating emotional speech with a concatenative synthesizer." In *ICSLP*, vol. 98, pp. 671-674. 1998.
[16]. Haojie Zhang; Yong Yang, "Fundamental frequency adjustment and formant transition based emotional speech synthesis," *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012*, vol.8, pp.1797-1801, 29-31 May 2012
[17]. Galanis, D.; Darsinos, V.; Kokkinakis, G., "Investigating emotional speech parameters for speech synthesis," in *proceedings of the Third IEEE International Conference on Electronics, Circuits, and Systems, 1996*, vol.2, no., pp.1227-1230 vol.2, 13-16 Oct 1996
[18]. Dr. Shaila D. Apte, "Digital Signal Processing", Second Edition, Wiley India.
[19]. Dr. Shaila D. Apte, "Speech and Audio Signal Processing", First Edition, Wiley India.