Abstract—This work proposes a method for synthesizing expressive speech from the given neutral speech. The neutral speech is processed by the Linear Prediction (LP) analysis to extract LP coefficients (LPCs) and LP residual. The LP residual is subjected to prosodic modification using the pitch, duration and amplitude parameters of the target expression. The LPCs of the neutral speech are replaced with that of the target expression using the Dynamic Time Warping (DTW). The synthesized speech using prosody modified LP residual and replaced LPCs sounds like the target expression speech. This can also be observed by the waveform, spectrogram and objective measures.

I. INTRODUCTION

Expression adds richness into the speech. For instance, we often enjoy happy expressive speech compared to sad expression. Expression also provides one more level of information which is non-linguistic. For instance, same linguistic message conveyed in two different forms of expression conveys different requirements in the conversation. A query about something asked in angry expression needs to be handled on a priority and careful basis than the same query asked in neutral expression. Human speech communication exploits the expressiveness available in speech to understand the level of involvement of listener in the ongoing conversation. Accordingly steps are taken to alter the conversation expression to increase the effectiveness of communication. Humans are good at producing and perceiving expressive information in speech. Apart from the linguistic message being conveyed, expressiveness plays a major role in effective communication. State of the art automatic speech processing systems mainly focus on extracting the linguistic information present in the speech signal. To increase expressiveness in human-computer interaction, methods need to be developed to extract and also incorporate the expressive information to the speech signal, termed more commonly as expressive speech synthesis.

There are several approaches proposed in the literature for the expressive speech synthesis [1] [2] [3] [4]. Most of these methods are based on concatenative speech synthesis framework which provide a high quality expressive speech [5] [6]. This is because the target expression speech signals for units at different levels are stored in the inventory and picked and concatenated for speech synthesis. For specific application where you need expressive speech from a predefined speaker for the required expression, expressive speech synthesized by concatenative synthesis framework is good enough. However, the limitation is the requirement of very large database storage and also limited options in terms of speaker and expressive quality. An alternative to concatenative expressive speech synthesis may be to store only neutral expressive speech and try to synthesize required expressive speech by modifying excitation source, vocal tract and modulation spectrum information. The present work is aimed in this direction.

One effective way of synthesizing the expressive speech is to modify the source and system features of neutral speech according to the parameters of the target expression. The neutral expression speech is processed by the Linear Prediction (LP) analysis to extract LP Coefficients (LPCs) and LP residual [7]. The LPCs mostly represent the vocal tract information and LP residual mostly represent the excitation source information. In the context of expressive speech, they also represent respective expressive components. The expressive information in the excitation source component may be characterized in terms of amplitude, pitch and duration. Thus if we have neutral expression residual, then it can be modified by using the amplitude, pitch and duration parameters of the target expression by prosody modification [8] [9]. The vocal tract component of neutral speech can be modified according to the target expression by mapping. One simple approach for the same is to use Dynamic Time Warping (DTW). The LPCs of neutral speech are mapped to LPCs of target expression speech by DTW. The selected target expression LPCs are representatives of the vocal tract information. Thus the synthesized speech using prosody modified LP residual and DTW mapped LPCs should sound like target expressive speech. Hence the motivation for the work.

The rest of the paper is organized as follows: The prosody modification of the LP residual using the instants of significant excitation is described in section II. The mapping of vocal tract spectral features of neutral to target expression spectral features using DTW is explained in section III. The experimental results and discussions are given in section IV. The summary of present work and scope for the future work are given in section V.

II. PROSODIC MODIFICATION OF NEUTRAL EXPRESSION LP RESIDUAL

Prosodic modification in conventional speech processing having only neutral speech includes pitch and duration modi-
Fig. 1: (a) LP residual of the neutral speech, (b) Instance of significant excitation, (c) Original epoch interval (o) and new epoch interval (*) for the prosody modification of pitch factor, $\alpha = 0.87$, duration factor, $\beta = 0.93$ and amplitude factor, $\gamma = 1.0$, (d) prosody modified residual and (e) LP residual of target angry expression.

The final residual is the one which represents the excitation source information for the target expression. Fig. 1 shows the LP residual of neutral speech, its instants of significant excitation, epoch interval plot, modified epoch interval plot and prosody modified LP residual. The LP residual of the actual target expression speech for the same text is also shown in figure. The prosodic manipulated LP residual resembles more like the actual target expressive speech LP residual in terms of associated pitch periodicity. Thus this LP residual may be able to act as necessary excitation signal for synthesizing target expressive speech.

III. DTW BASED MAPPING OF LPCS OF NEUTRAL SPEECH

The next step in the expressive speech synthesis is to modify the LPCs representing the vocal tract information. One approach is to find the mapping from the neutral LPCs to the target expression LPCs. Accordingly, LPCs of neutral expression can be mapped to the target expression. Alternatively if we have access to the target expression LPCs, then we can find the best matching LPCs for synthesis by using DTW. This process works only for the text dependent case. The LPCs of neutral and expressive speech signals are converted to LPCCs. The LPCCs are matched by the DTW technique based on least distance cost [10]. The pairs which lead to total least distortion give frame indices of target expression speaker LPCs. The matched LPCs of target expression are used as the time varying filter parameters for synthesizing speech.

Fig. 2 shows the LP spectrum of frame of neutral speech, speech synthesized by using neutral residual and matched LPCs from the target expression. The LP spectrum of a frame of target expression speech is also plotted in the figure. The synthesized speech using the matched LPCs seem to be more like target expressive speech in terms of LP spectra. This can be observed from the Fig. 2 (e and f).
IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

The expressive speech synthesis can be accomplished by exciting the time-varying filter whose parameters (LPCs) are obtained by DTW matching using the prosodic modified LP residual. Speech signals for the same text are collected from two male and two female speakers for the neutral, happy and angry expressions. The speech signals are sampled using 8 kHz and stored with 16 bits/sample resolution. All the speech signals are processed by the LP analysis to extract LP residual and LPCs. The LP residual of neutral speech is subjected to prosodic modification using the pitch, duration and amplitude factors derived from the expressive speech. The LPCs of neutral expression speech are matched with LPCs of expressive speech using DTW technique after converting them into LPCCs. After matching, the respective frame LPCs are chosen from the target expression for synthesizing the speech. Speech signals are synthesized using the prosodic modified version of the LP residual and matching LPCs chosen by DTW.

Speech signals of neutral expression, synthesized angry expression and actual angry expression and their narrow band spectrogram are shown in Fig. 3. As it can be observed, the synthesized speech signal looks more like angry expressive speech both in the temporal as well as spectro-temporal domains. It should be noted that the actual angry expression speech is a different example from the same speaker. Speech signals synthesized from the female speakers for happy emotion are shown in Fig. 4. In this case also, the synthesized speech looks like happy emotion expressive speech.

To find the effectiveness in incorporating the expressiveness objectively, the LPCs of synthesized expressive speech are compared with the neutral expression and target expression speech signals by DTW and the total cost in comparison are obtained. The total cost for angry-angry comparison (intra expression) is 0.17 compared to 0.27 of angry-neutral expression (inter expression) comparison. The total cost for intra expression comparison is less compared to inter expression comparison. This infers that the synthesized speech by LP residual modification and LPCs matching indeed produces expressive speech in the target expression.

The present work is a simple illustration of the possibility of the expressive speech synthesis using LP residual modification and LPCs matching. There are several shortcomings associated with the present work. First, the prosodic parameters of the target expressions are obtained from only one example. Hence the generality of the result is to be questioned. To alleviate this, a large database under different expressions needs to be collected and analysis should be done to obtain statistically significant and reliable values for prosodic parameters like pitch, duration and amplitude factors. Secondly, pitch, duration and amplitude are only estimated at gross level. However, expressive information is present at fine levels also. Hence careful analysis is required to know the variations of the excitation source information at the finer level. Methods may then be developed to estimate the parameters representing them from speech and use them for expressive speech synthesis. Coming to the vocal tract information part, currently, LPCs of target expression are obtained by DTW matching. This is a trivial
assumption. However, in practice LPCs of target expression may not be available. To handle this, what is to be done is to develop a mapping of LPCs from neutral to target expression. Of course, the mapping should be sound dependent. Once this mapping is available, then these parameters may be used to alter the LPCs of neutral speech to obtain the LPCs of the target expression. Finally, there is a distortion introduced into the synthesized expression both due to prosodic manipulation and different set of LPCs. The perceivable distortion hampers the appreciation of the synthesized target expression. This is because both naturalness as well as expressiveness are such fine features, any slight alternation due to signal manipulation leads to significant reduction in the perception quality of these factors. New methods for speech signal manipulation are required.

V. SUMMARY AND CONCLUSIONS

This work proposed a method for expressive speech synthesis by LP residual prosodic modification and LPCs DTW matching using the observed target expressive prosody parameters. The synthesized speech sounds like a target expression speech. This is also observed in the waveforms, spectra, spectrograms as well as objective measures.

In the present work prosody parameters (α, β and γ) are derived by manual observations of the target expressive speech. Future work should focus on estimating the reliable and statistically significant prosodic parameters. The work should also focus on extracting the fine level excitation source features for expressive speech synthesis. Also the mapping needs to be developed for vocal tract features. Finally improved speech signal processing methods which introduce least distortion should be developed.

REFERENCES