Improving intelligibility of synthesized speech in noise with emphasized prosody

Sunil R. Shukla
Department Electrcial and Computer Engineering, Georgia Institute of Technology, USA

https://doi.org/10.36505/ExLing-2010/03/0041/000161

Abstract

The performance of current high quality concatenative text-to-speech (TTS) systems is limited under noisy environments. This paper investigates whether or not the intelligibility of synthesized speech in noise can be improved by emphasizing the prosody. Additionally, the paper presents a method that can effectively emphasize the prosody of units in existing TTS databases. The circular linear prediction (CLP) model is combined with the constant-pitch transform (CPT) to perform pitch and duration modifications to concatenative TTS units with little impact to the subjective quality. Test utterances are generated using the method and compared to reference utterances synthesized by a high quality TTS engine. The subjective test results demonstrate a preference for emphasized prosody in the majority of the test cases.

Key words: TTS, speech synthesis, linear prediction, prosody, noisy speech.

Background

Spoken dialogue interfaces have become highly common in numerous applications including telecommunications, vehicle navigation, and vehicle command and control. These applications are often used in environments that are prone to significant background noise. It is well-known that noisy environments can significantly impact the intelligibility of synthesized speech (Langner and Black 2005). Even the better performing of the currently available text-to-speech (TTS) products are, on average, 22% less intelligible than human voice under comparable noisy conditions (Venkatagiri 2003). This study explores the hypothesis that a system capable of adding emphasis to key words or syllables in the utterances can improve intelligibility. The current top performing high quality, concatenative TTS systems avoid signal processing for synthesis. Prosody realization, for these systems, is achieved by selecting the closest matching units from a large, prosodically rich database. Hence, emphasized prosody realization would require a significantly larger database. This study presents a novel approach for realizing emphasized prosody, and investigates its efficacy in improving intelligibility of concatenative TTS in “real world” noisy environments.

The CLP/CPT Representation

Although current speech models are effective, and highly useful (especially for speech coding), they are inherently inaccurate due to the quasi-stationary characteristic of speech signals. Current TTS systems are motivated to avoid signal processing for prosody realization due to the potential for audible artefacts caused by these modelling inaccuracies.

Circular linear prediction (CLP) is a highly accurate parametric model for speech (Shukla et al 2002). It improves upon classic LPC techniques (i.e. residual-excited LP) by providing a windowless method for calculating the LP coefficients while retaining the stable filter coefficients property. This reduces the modelling errors attributed to applying a window to each analysis frame and overlap-add during synthesis. In this method, each LP analysis frame is exactly the length of the local pitch period. In this case, each pitch period can be analyzed as an infinitely periodic signal. As in other LP-based methods, the parameters for the segment database are the residual signal, the CLP coefficients, and the pitch period of each frame.

The effectiveness of the model is based on the assumption that the exact pitch period of voiced speech can be detected. This is satisfied by performing the analysis with fractional resolution by oversampling the signal. Note that for unvoiced speech, the assumption becomes irrelevant. In practice, it was found that an oversampling factor of 10 provides sufficient resolution for accurate modelling without audible artefacts (Ertan 2004). Pitch period detection of the oversampled signal is conducted iteratively, by choosing the frame length that maximizes the CLP gain. The complexity of this algorithm is not an issue since analysis for TTS is not conducted in "real-time".

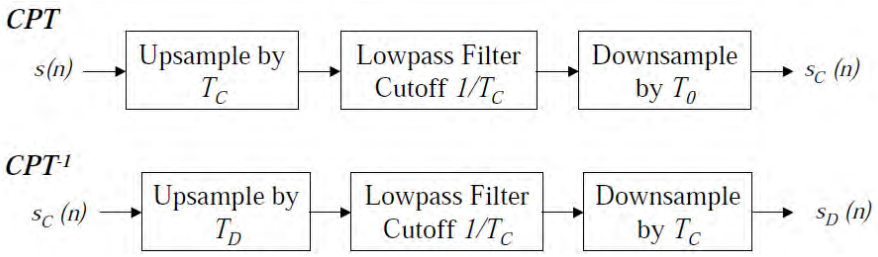

Figure 1: Block diagrams of the constant-pitch transform and the inverse.

The constant-pitch transform (CPT), as shown in figure 1, is a method for modifying the period, T_0, of every frame of speech, $s(n)$, to a constant length, T_C (Shukla et al 2002). Following CLP analysis, the CPT is applied to every frame to result in a TTS segment database with uniform pitch. Effectively, this technique removes phase discontinuities, which aids in reducing the complexity of segment concatenation and pitch modifications. The

CLP/CPT representation provides a novel technique for realizing parametric modifications to speech segments during synthesis, while having little effect on the synthesis quality.

Prosody Modification and Synthesis

The CLP/CPT representation reduces the relative complexity of prosody modification and segment concatenation, since all units are phase-matched. Modifying the frames segment to a desired pitch, T_D, is achieved by the inverse CPT (see figure 1). Phoneme durations are modified by repeating or deleting frames and parameters. Segment concatenation does not require overlap-add or interpolation, assuming that the database has been normalized. Note that voiced/unvoiced transition regions require additional considerations, which are detailed by Shukla and Barnwell (2007).

Emphasized Prosody

To emphasize the prosody of key words in an utterance, this paper assumes that the locations of key words of the utterance have been marked. In this case, a simple algorithm has been employed for adding emphasis:

(1) increase the target pitch and duration of the voiced phoneme of the primary syllable by 10-12%;
(2) lower the pitch of the voiced phonemes of the adjacent syllables by 5-7%.

The above algorithm is applied to the prosody curve generated by an existing high quality, unit-selection synthesis TTS engine.

Subjective Testing

To compare the speech intelligibility improvements, this study implements the CMU Communicator, a spoken dialogue interface that includes a high quality, limited domain, concatenative TTS engine (Rudnicky et al 1999). A subjective test was designed by creating a simulated conversation between the CMU Communicator system and a user in a noisy environment such that the intelligibility of the reference synthesized utterances is degraded.

Method

The CLP/CPT analysis method was applied to the CMU Communicator synthesis database. The test utterances were generated by re-synthesizing the reference utterances with the key words emphasized. Road and traffic noise at highway speeds was recorded and added to the synthesized reference and test utterances. The test is set up such that the subjects are observers to a conversation between a user and the CMU Communicator. The user requests the synthesizer to repeat the response. The subjects then select whether they would prefer the CMU Communicator to repeat the utterance with the same or emphasized prosody. A total of 19 sets of

reference and test utterances were synthesized, and each set were presented to the subjects in random order.

Results and Analysis

Of the 285 total selections made by 15 different subjects, the emphasized responses were preferred 59% of the time, with high statistical significance. The confidence interval for the preference was greater than 99.5%. For some of the cases in which the unmodified speech was preferred, it was noted that the emphasized speech contained audible artefacts. Additionally, in certain cases, the test utterances sounded either overemphasized or not having enough emphasis. These issues are attributed to a suboptimal algorithm for determining the target prosody values and a limitation to the extent of modifications the CLP/CPT method can achieve.

This study concludes that the emphasized prosody is generally preferred in noisy environments. An effective method for realizing emphasized prosody for existing unit-selection TTS synthesizers is presented. The method can be utilized to either modify the prosody during synthesis, or increase the richness of the uni-selection database. During implementation, subjective tests revealed that the range of modifications for the CLP/CPT method is limited to 10% - 15%.

References

Ertan, A.E., Shukla, S., Barnwell, T. 2002. Circular LPC modeling and constant pitch transform for accurate speech analysis and high quality speech synthesis. ICASSP, 269-272, Orlando, USA.

Ertan, A.E. 2004. Pitch-synchronous processing of speech. Georigia Institute of Technology: School of Electrical and Computer Engineering, Ph.D. Thesis.

Langner, B., Black, A. 2005. Improving the understandability of speech synthesis by modeling speech in noise. ICASSP, IV-265-268, Philadelphia, USA.

Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu W., Oh, A. 1999. Creating natural dialogs in the Carnegie Mellon Communicator system. Proc. of Eurospeech, vol. 4, 1531-1534, Budapest, Hungary.

Shukla, S., Barnwell, T. 2007. Improving high quality TTS using circular linear prediction and constant pitch transform. ICASSP, 681-684, Honolulu, USA.

Venkatagiri, H.S. 2003. Segmental intelligibility of four currently used text-to-speech synthesis methods. The Journal of the Acoustical Society of America, 113, 2095-2104.