METHODS FOR IMPROVING THE QUALITY OF SYLLABLE BASED SPEECH SYNTHESIS

Venugopalakrishna Y R [1], *Vinodh M V* [2], *Hema A Murthy* [2], *C S Ramalingam* [1]

[1]Department of Electrical Engineering,
[2]Department of Computer Science and Engineering,
Indian Institute of Technology Madras,
Chennai 600 036

ABSTRACT

Our earlier work [1] on speech synthesis has shown that syllables can produce reasonably natural quality speech. Nevertheless, audible artifacts are present due to discontinuities in pitch, energy, and formant trajectories at the joining point of the units. In this paper, we present some minimal signal modification techniques for reducing these artifacts.

Index Terms— Speech Synthesis, Syllable, Prosody, Silence correction, Stop sounds, LSF interpolation

1. INTRODUCTION

Most of today's commercial TTS systems employ unit selection based concatenative synthesis. In unit selection based speech synthesis [2], the goal is to pick the best unit sequence from a pre-recorded continuous speech database to match the given target unit sequence. The speech database is designed such that each unit is available in various prosodic and phonetic contexts. We can consider the speech database as a state transition network with each unit in the database occupying a separate state. The state occupancy cost (target cost) is the distance between a database unit and a target unit, and the transition cost (join cost) is an estimate of the quality of concatenation of two consecutive units. A pruned Viterbi search is used for selecting the best unit sequence with the lowest overall cost (weighted sum of target cost and join cost).

Well-known unit selection based speech synthesizers such as Festival [3], AT&T Next-Gen [4], etc. use smaller units (e.g., diphones, phonemes, etc.) for building voices in various languages. Thomas et al. in [1], and Kishore and Black in [5] have shown that syllable-like units are a good choice for synthesis in Indian languages. In [6] we described the design and implementation of a unit selection based text-to-speech synthesizer with syllables and polysyllables as the basic units. Synthesizers for two Indian languages (Hindi and Tamil) were built for IVR applications. The Hindi synthesizer was built using two hours of data, whereas the Tamil synthesizer was based on a one-hour database. The synthesized speech quality was clearly natural sounding, but there were audible artifacts causing a drop in the perceived overall quality.

In this work we have attempted to understand the causes for these artifacts and present methods for improving the quality of the synthesized speech by applying minimal prosodic modifications. This paper is organized as follows: In section 2 we discuss the methods for minimizing artifacts due to target mismatch, energy discontinuity, incorrect silence for stop sounds, and spectral discontinuity at joining points. Section 3 gives a brief overview of our focus for the future.

Thanks to DST (project no. ELE0506118DSTXCSRA) and IBM (project no. CSE/06-07/087/IBMC/HEMA) for funding.

2. METHODS TO MINIMIZE ARTIFACTS

2.1. Corpus used

A Hindi language database of two hours duration, recorded by a voice talent in an anechoic chamber, is used for synthesis. It comprises of 2360 sentences. These recorded sentences were segmented into syllable sized units using group-delay based segmentation technique [7]. The segments were then labeled manually. This resulted in a total of 66552 units of which 8701 were unique.

2.2. Classification of syllable-like units

Temporal and spectral shape of the syllable units vary based on their position in the word and in the sentence. For example, syllable units at the end of a word commonly have a falling temporal energy pattern. Hence, the position of the syllable in the word (begin, middle and end), and position of the syllable in the sentence are important target features. In this work, we have classified syllable units in the database according to their position in the word as begin<syl>, mid<syl>, and end<syl>. During unit selection, the units are picked based on this classification. Begin<syl> corresponds to unit <syl> obtained from the beginning of a word, mid<syl> corresponds to unit <syl> obtained from the middle of a word and end<syl> corresponds to unit <syl> obtained from the end of a word.

2.3. Energy Prediction and Modification

Even with careful recording of prompts, the intensity with which a voice talent reads the prompt varies over the length of the recording. In addition, syllable-like units used in concatenation are picked from different contexts. Because of these reasons, audible discontinuity due to energy mismatch is present in the synthesized speech. To counter this, we modified the energy of the selected units based on Classification and Regression Tree (CART) prediction.

For building the CART model, the identities of the current, previous and next units, position of the syllable in the word (begin, end, middle, single) and in the sentence are used as input features, whereas peak amplitude level (measure of energy) of the syllable unit is used as the output feature. CART was built from a data set made up of 1180 sentences corresponding to 1 hour of speech. The correlation coefficient was 0.91.

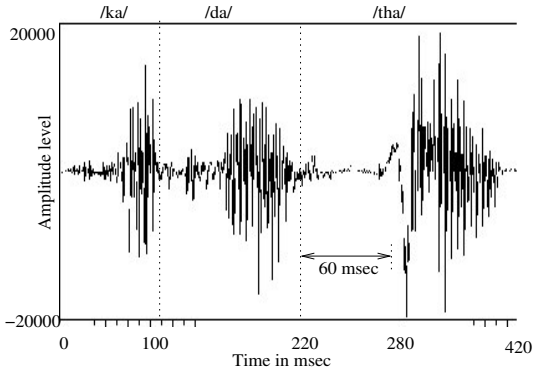
This CART tree is used for predicting the peak amplitude of each syllable in a sentence. The selected syllable units are then scaled appropriately to match the predicted peak value. MOS test results showed an improvement of 0.4 after energy modification.

2.4. Silence correction for stop sounds

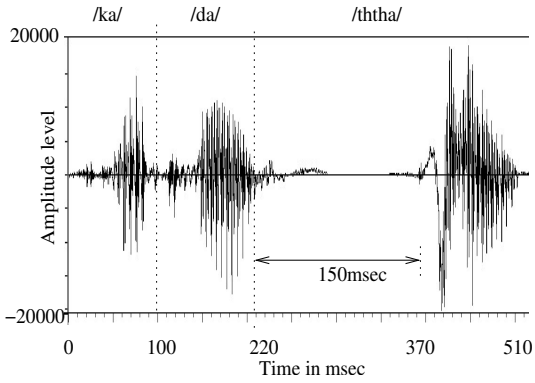
The duration of the closure portion of a plosive sound plays an important role in the quality of synthesis. For plosives, because units are picked from different contexts and partly because of faulty labeling of silence, concatenating them results in a duration of silence that may not be appropriate for the target context. More importantly, this duration is the one that distinguishes gemination and non-gemination (i.e., long and short consonants). Usually, silence duration for geminates is between 2 to 2.5 times that of non-geminates. Hence, inappropriate duration of silence will lead to plosives sounding unnatural, or a geminate being perceived as a non-geminate (and vice versa) . For example, on synthesizing a Tamil word /kadaththa/, silence duration for geminate /thth/ was 60 msec and sounded like /kadatha/, and not natural. However, on changing the silence duration to 150 msec it not only sounded like /kadaththa/, but the quality also improved. Fig. 1(a) and Fig. 1(b) show the stop duration in /kadatha/ and /kadaththa/ respectively.

Observation shows that, if an onset stop appearing before a coda of a closed syllable has less silence than required, the coda was not perceivable. For example, in case of "ek kauvaa", if the silence of onset stop /k/ of "kau" was much less than 50msec, the coda /k/ of "ek" was not perceivable.

To reduce the above artefact, a silence detection and correction module was introduced in the synthesizer engine. The silence detection module estimates the amount of silence for each selected unit using a simple energy based silence estimation algorithm. Subsequently, the silence of the unit is corrected based on the silence duration of onset stops obtained through the following analysis.



(a) Waveform showing /kadatha/



(b) Waveform showing /kadaththa/

Fig. 1. Waveforms showing significance of silence duration of stop sounds

A study on silence for all possible open syllables, geminates and for stops appearing after the coda of a closed syllable was done. The average duration of silence for various stops is given in Table 1. For each stop, the silence duration is measured from randomly chosen hundred examples, fifty for units appearing in the beginning position of a word and fifty for middle or ending position of the word. Some of the important observations are mentioned below.

- To study the variation of the silence duration of onset stop consonants due to different nuclei, the consonant /k/ as onset with various short vowels /a/, /i/, /u/, /e/, /o/, /au/ as nucleus was considered. As shown in Table 1, the average silence duration is around 50 msec with ± 8 msec tolerence. From this, we can conclude that the silence duration of onset stops do not show drastic variation due to different nuclei of syllables.

- Syllables “ka”, “ki” and their aspirated counterparts “kha”, and “khi” were chosen to study the effect of aspiration the amount of silence. The average silence duration was around 50 msec.
- As we ensured that the amount of silence of onset stops do not vary drastically with different nuclei, fifty examples were chosen with various nuclei for all other stop sounds /ch/, /T/, /t/ and /p/. Out of all these /ch/ had 26 msec, which is least and all other sounds /T/, /t/ and /p/ had an average around 50 msec.
- For geminates “kk”, “chch”, “TT”, “tt”, and “pp”, silence duration is 2 to 2.5 times more compared to their non-geminate counterparts.
- For stops appearing after the coda of closed syllables, silence duration is 1.5 to 2.5 times more compared to that of stops appearing after open syllables. If coda of a preceding syllable is /k/, stops show increase in silence by about 1.5 times, whereas for the coda /p/ increase in silence of onset of next syllable is 2 to 2.5 times.
- Amount of silence is nearly same for syllable appearing in begin and end positions.

Stop	avg. silence (msec)		Stop	avg. silence (msec)	
	begin	mid/end		begin	mid/end
ka	52.8	42.3	pa	52.2	53.8
ki	43.7	40.3	ba	52.8	49.2
ku	50.6	51.0	chch	–	69.8
ke	46.6	48.4	kT	–	79.0
ko	56.2	46.8	TT	–	80.0
kaa	58.0	51.0	phT	–	70.0
kki	–	114.8	kD	–	69.0
kha	59.0	43.2	kt	–	72.0
khi	52.6	*	cht	–	35.0
ga	29.6	33.2	tt	–	78.6
cha	26.4	34.2	dd	–	70.0
ja	31.6	*	pk	–	105.0
Ta	46.8	39.4	pch	–	84.0
Da	*	31.4	pT	–	87.0
ta	41.2	43.0	pt	–	120.0
da	46.4	32.6	pp	–	106.0

* indicates unit is not available in the database
– indicates unit may not appear in that position

Table 1. Duration of silence for onset stops of syllables appearing in begin and mid/end position of words

2.5. Smoothing Techniques

In unit selection synthesis, as units cannot appear in all possible contexts, adjoining units may not have smooth tranisition in formants at their joining points. Past studies have shown that smooth changes in frequency are perceived as changes within a single speaker, whereas sudden changes are being perceived as a change in speaker [8].

To reduce the effect of audible formant transitions, spectral smoothing and spectral interpolation techniques are employed in [9]. In that work, various techniques such as optimal coupling, waveform interpolation, LP pole shifting, LSF based smoothing and shaped noise methods (closure) were studied. It was concluded that no single technique was superior.

2.5.1. Line Spectral Frequency (LSF) based smoothing

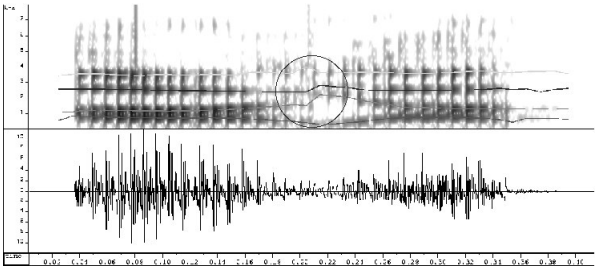
Linear prediction based models allow us to work on the system spectrum. If there is a discontinuity in the spectrum at the joining point of two units, the smoothing is usually done in the Line Spectral Frequency (LSF) domain [10]. This well-known procedure is followed because stable filters continue to remain stable even after linear interpolation.

The frames used for linear interpolation are the last frame of the first segment and the first frame of the succeeding segment. These frames are called anchor frames. Pitch synchronous frames of two pitch periods are considered, i.e., first anchor frame starts from the last but two pitch mark sample and ends at the last pitch mark sample of the first segment, and the second anchor frame starts from the first pitch mark sample and ends at the third pitch mark sample of the second segment. LP analysis is performed on both the frames to derive the LPCs and the residual. LPCs are converted in to LSFs and then LSF zeros are linearly interpolated to get the LSFs for the new frames. Whereas, the residual for the new frames is obtained by waveform interpolation of the residual of the anchor frames. LSFs of the new frames are converted back to LPCs and are used in the synthesis filter with waveform interpolated residual as excitation to obtain new speech frames. These new speech frames are inserted at the joining point. This was carried out by varying the number of new frames to be inserted. Two frame insertions showed good results. Formant plots for the sound “aayaa”, before and after LSF based interpolation at the join of units “aa” and “yaa”, are shown in Figure 2.

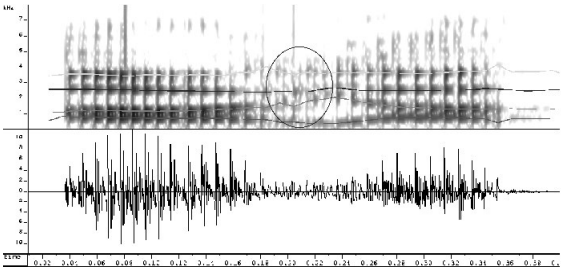
3. ISSUES TO BE ADDRESSED

3.1. Co-articulation due to geminates

Since syllables, rather than phonemes, are used for concatenation, the artifacts due to lack of co-articulation will be less, as we will have less number of joins and the joining boundaries are low energy points. However, segmented units occuring just before geminates will still be left with some co-articulation due to the following geminate. For example, consider the Tamil word “muppaththu” being segmented and labeled as “mu”, “ppa”, and “ththu”. At the end of the seg-



(a) Formant plot before interpolation



(b) Formant plot after interpolation

Fig. 2. Formant plots for sound "aayaa" before and after LSF based interpolation at join of units "aa" and "yaa"

mented unit "mu" there will be some co-articulation corresponding to /p/ due to lip closure foreseeing the geminate "ppa". If this unit is used for a target "mu", there will be a significant mismatch. Therefore, there is a need to handle such co-articulation related issues in syllable based synthesis.

3.2. Small footprint synthesizer

Another important issue is to reduce the overall footprint of the synthesizer while improving its quality.

One possible approach, is to create a mapping mechanism that can transform the signal parameters (like Pitch, Energy, Duration, etc.) of a reference syllable to the that of a contextually different target syllable. Here, it would be sufficient to store only a few representative waveform(s) of a syllable, and model the signal parameters for other contextually different target syllables.

Another approach is to build a syllable based parametric speech synthesis system using the HTS (HMM based) engine [11]. Existing HMM based synthesizers are based on phonemes or diphones, whereas this effort aims at building generative HMMs for syllables and to generate speech parameters from them. As these two approaches are parametric, they help in reducing footprint of the system.

4. CONCLUSION

We have proposed to match the prosodically rich syllable units to the target by classifying them according to their position in the word. This is based on a study that syllables in different positions of a word sound different. In addition, we have modified energy of syllables based on CART prediction of their peak syllable amplitude. Correction of silence for onset stops helps in reducing the artifacts introduced by faulty segmentation and labeling. LSF based smoothing of joins aids in removing the audible artifacts by improving the continuity in formant trajectory.

5. REFERENCES

[1] Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy and C. S. Ramalingam, "Natural Sounding TTS based on Syllable-like Units," in *European Signal Processing Conference*, Florence, Italy, 2006.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of IEEE ICASSP*, 1996, vol. 1, pp. 373–376.

[3] CSTR, "Festival Speech Synthesis System," http://www.cstr.ed.ac.uk/projects/festival/, 2003.

[4] AT&T Research Labs, "The AT&T Next-Gen TTS System," http://www.research.att.com/projects/tts/, 1999.

[5] S.P. Kishore and A.W. Black, "Unit size in unit selection speech synthesis," in *Proceedings of EUROSPEECH*, 2003, pp. 1317–1320.

[6] Venugopalakrishna Y R, Sree Hari Krishnan P, Samuel Thomas, Kartik Bommepally, Karthik Jayanthi, Hemant Raghavan, Suket Murarka and Hema A Murthy, "Design and Development of a Text-To-Speech Synthesizer for Indian Languages," in *NCC*, 2008, pp. 329–331.

[7] T. Nagarajan, H.A. Murthy and, R.M. Hegde, "Segmentation of speech into syllable-like units," in *Proceedings of EUROSPEECH*, 2003, pp. 2893–2896.

[8] B.C.J Moore, *An Introduction to the Psychology of Hearing*, Academic Press, New York, 1997.

[9] D.T. Chappell and J.H.L. Hansen, "Spectral Smoothing for Speech Segment Concatenation," *Speech Communication*, vol. 36, pp. 3–4/343–373, 2002.

[10] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, pp. 35, 1975.

[11] Nagoya Institute of Technology, "HTS," http://hts.sp.nitech.ac.jp, 2002.