Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 3– No.7, December 2015 – www.caeaccess.org

Marathi Synthetic Voice using Synthesizer Modules of Festival Speech and HTS Straight Processing

Sangramsing N. Kayte
Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

Charansing N. Kayte
Digital and Cyber Forensic, Aurangabad, Maharashtra, India

Bharti W. Gawali
Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India.

ABSTRACT

A new Maharashtra Marathi voice was created using four festival modules Clunits, Clustergen, Multisyn and HTS, as well as an additional database was created with straight processing. All voices were created using the same database to allow for consistency and for easier comparison of the output. Once these voices have been created they can be used as a baseline for further development in Maharashtra Marathi speech synthesis. Except for the Multisyn module, which has some problems due to coverage, results are acceptable, with the HMM designed voices having the best quality.

Keywords

Speech synthesis, straight, festival, HTS.

1. INTRODUCTION

The new speech synthesis techniques: hts and straight, have improved the recent research on this field. Specifically, looking for natural speech the HTS-Straight technique has had excellent results. We tried to check these results for Maharashtra Marathi Language. However, the HTS technique is easy to insert in the well know Festival System, with some modification to the system to account for missing parameters, so it is easy to compare HTS with traditional modules of Festival as Clunits and Clustergen [1]. In this article, we compared the quality of synthesized speech of Festival modules Clunits, Clustergen, Multisyn and HTS for Maharashtra Marathi. The use of these modules and techniques requires not only a straightforward adaption to Maharashtra Marathi. In some cases, especially for the HTS-Straight Technique, the adaption required a strong software redesign. Other purpose of the article was to check the buzzy effect of the HTS Module for Maharashtra Marathi, more carefully than we have done before [2][6][7].

2. BASIC DESCRIPTION OF SYNTHETIZER MODULES

2.1 Clunits and Multisyn Modules

The Clunits (Cluster Units) method works by extracting a list of phones from a set of prerecorded phrases, including their prosodic context, generating a Classification and Regression Tree, that according to context will give a set of possible segments to use depending on context at synthesis time. During synthesis, a target is generated for each phone to be synthesized based on its context. Once a set of possible units is extracted from the Classification and Regression Tree for each target, an optimal path is generated, concatenating the units that will generate the smallest weight throughout the phrase. Multisyn works in a similar fashion, replacing phone selection with di-phones [8][9][10][11].

2.2 HTS and Clustergen Modules

The HTS module is based on Hidden Markov Models. These HMMs are used to generate the decision trees used to select the optimal set of parameters during synthesis time. The HTS module uses three sets of parameters (Mel Cepstral, F0 and phone duration). Each set of parameters is extracted from the database independently from the others, allowing for prosody modifications at the cost of some distortion due to the source/filter model used during synthesis time. Clustergen works the same as HTS, with only some modifications on the way the HMMs are generated [12][13][14][15].

2.3 Straight

Straight synthesis is done using HTS, by replacing the Mel-Cepstral parameters with Straight parameters. However, the HTS module in festival does not allow for the use of Straight parameters. In this case, Festival is used exclusively for prosodic analysis, using this output to feed a set of external applications for parameter selection and synthesis

3. DATABASE RECORDING AND LABELING

The database used for the voice generation consists of approximately 90 minutes of poetry. Each phrase around 1000 was stored in a separate file. A transcription file was also generated to be used for the phone labeling of each audio file. Additionally, a set of Festival SCHEMA files were created. These files contain the rules for extracting the phones from the text input, including some exceptions and handling of numbers, dates and basic formatting. The Schema files were adapted from 2 existing Castilian voices, using a modified phone-set to better model the Mexican variant of Spanish, and we adopted small changes not used before [2]. For this database 66 phones were used 38 consonants and 28 vowels, see Table I. Each of the consonant phones is classified according to three categories, while the vowels were classified in four categories. In the case of all HMM based voices, the labeling was carried out with the EHMM labeler, using the same labeling in all cases. The labeler outputs the starting, middle and ending point of each recorded phone. In all cases some manual verification and correction was carried out [3][16][17][18].

- Consonant Classification By Type:
- S (occlusive): The oral and nasal cavities are closed, there is no air flow.

- F (fricative): There is constant friction in the articulation point, the air flow is not completely restricted.
- A (affricative): Formed by an occlusive sound, followed by a fricative.
- N (nasal): The oral cavity is closed, air flows through the nose.

3.1 Vowel Classification

By Height: the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw (Low, Mid or High).

By Position: the position of the tongue during the articulation of a vowel relative to the back of the mouth (Front, Mid or Back).

Lip Rounding: The lips are (+) or not (-) in a rounded position.

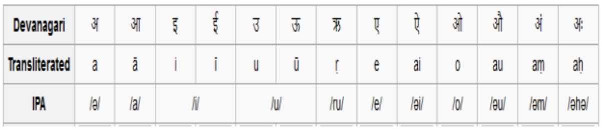
Stressed: The vowel is (+) or not (-) stressed.

4. TRAINING PROCESS

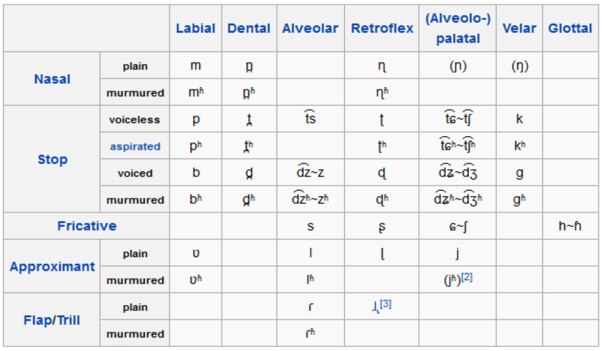
4.1 Festival Training

The three festival based voices Clustergen and Clunits were trained using the Festvox software. Clunits [5]: For clunits, extracts all instances of each phone and clusters them according to their context. During synthesis time, the context is extracted and the corresponding cluster is extracted, selecting the set of segments with smaller cost based on their Cepstral parameters. Multisyn For multisyn, diphones are used, selecting an optimal path at synthesis time. If a diphone can't be found, a back off module is used to replace with an appropriate replacement. Clustergen Clustergen training is HMM based, creating a set of CART trees (Mel Cepstral, F0 and duration), each set of parameters calculated independently from each other. Clustergen is phone based, clustering phones according to their context [4], [5] [19][20][21][22].

Table 1: Phones Used in the voice bank



Devanagari	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः
Transliterated	a	ā	i	ī	u	ū	ṛ	e	ai	o	au	aṃ	aḥ
IPA	/ə/	/a/	/i/		/u/		/ru/	/e/	/əi/	/o/	/əu/	/əm/	/əhə/



		Labial	Dental	Alveolar	Retroflex	(Alveolo-) palatal	Velar	Glottal
Nasal	plain	m	n̪		ɳ	(ɲ)	(ŋ)	
	murmured	mʱ	n̪ʱ		ɳʱ			
Stop	voiceless	p	t̪	t͡s	ʈ	t͡ɕ~t͡ʃ	k	
	aspirated	pʰ	t̪ʰ		ʈʰ	t͡ɕʰ~t͡ʃʰ	kʰ	
	voiced	b	d̪	d͡z~z	ɖ	d͡ʑ~d͡ʒ	g	
	murmured	bʱ	d̪ʱ	d͡zʱ~zʱ	ɖʱ	d͡ʑʱ~d͡ʒʱ	gʱ	
Fricative				s	ʂ	ɕ~ʃ		h~ɦ
Approximant	plain	ʋ		l	ɭ	j		
	murmured	ʋʱ		lʱ		(jʱ)[2]		
Flap/Trill	plain			ɾ	ɽ[3]			
	murmured			ɾʱ				

4.2 HTS Training.

For HTS training while labeling was done using the EHMM tool provided with festival, training itself is used the HTK tools. For HTK training, a set of „questions‟ must be provided, that will contain the context information that will be used for the creation of the decision tree. This information must match the context information generated by the Festival labeling, but manual adjustments can be done. The HTK/HTS tools also provide a set of parameter that allows easy modification of the parameterization of the audio data. Due to this fact, different HTS voices were created to validate the effect of different factors in the quality of the synthetic voice [23][24][25].

4.3 HTS Parameters

Frequency Warping: This parameter allows the use of Cepstral or Mel-Cepstral parameters. Voices were created using no frequency warping and Mel Scale. As we expected, better results were obtained when using the Mel warped parameters. Gain: Log or Linear gain. Notice that the festival HTS module must be modified for log gain to work from festival. Log gain provides slightly better results, but at the cost of a much higher training time [26][27].

Gamma: This parameter affects the reconstruction filter parameters placement of poles and zeroes. Values of -1,-1/3 and 0 were used. Best results were obtained using -1/3, but care must be taken with the number of cepstral parameters, as the filter can become unstable. Number of Cepstral Coefficients: Vectors consisting of 12, 24 and 36 parameters were used. With 12 coefficients the reconstructed signal is too distorted. With more coefficients the reconstructed signal is clearer, but with a high number of coefficients the filter can become unstable, see Fig. 1 and 2 [28][29].

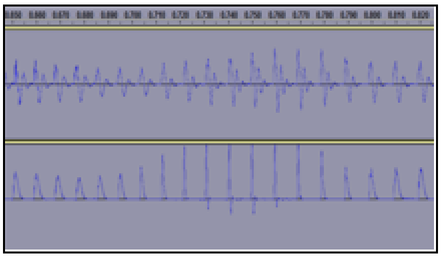


Fig. 1. Signal with 36 and 12 coefficients.

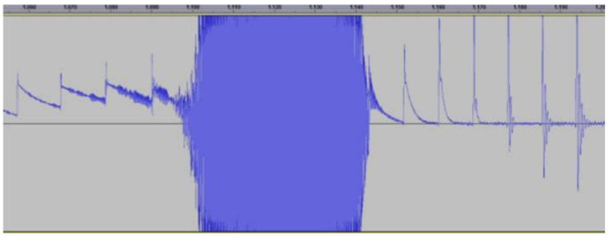
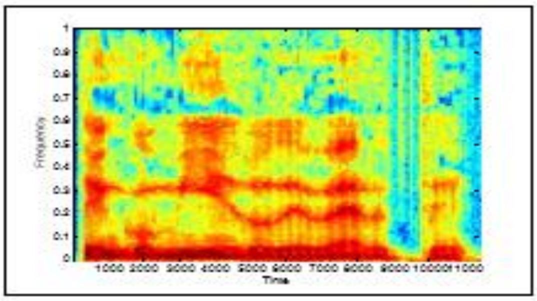


Fig. 2. Signal with 36 coefficients and Gamma=-1/3.

Number of states per phone: Each phone is divided into a number of HMM states, see Fig. 3. Voices were created with 3, 5 and 7 states.



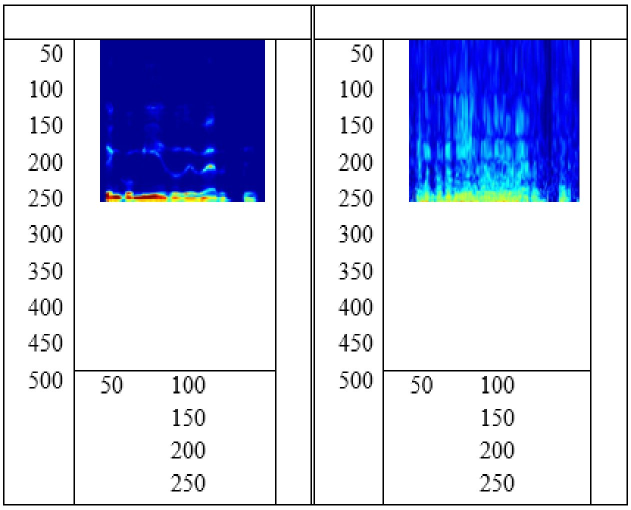


Fig. 3. a) Original spectrum. b) Smoothed spectrum. c) Aperiodicity data

5. RESULTS

The Marathi voices synthesized with the concatenative approach were found to have discontinuities at synthesis time. Of particular notice is the Multisyn module, as the Marathi database used was not phonetically balanced, resulting in high discontinuities and gaps in the synthesized speech. As the other modules are phone-based, they avoid this problem. The voices synthesized by HTS and HTS-Straight were valuated as very natural by four linguistic experts in our lab, more the second one. We did not used a MOS test, we preferred experts to check fundamental features of the voice. The buzzy effect is not relevant for Maharashtra Marathi, three of four experts did not heard it, and the only one heard “something unnatural” in the hts technique but not defined it as buzzy voice. We designed an interface for easy use of these techniques, and promptly will be ready for free internet access. From 4 to 6 states per phone there is marked improvement. Beyond 6 states, this is minimal. And the experts split decisions about quality rise of the voice from 5 to 8 states. Finally, we hope these experiences will help researchers for Spanish Language in the use of these techniques.

6. REFERENCES

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. of IEEE ICASSP, 2000, pp. 1315–1318.

[2] Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015

[3] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte “Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language” OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 5, Ver. I (Sep –Oct. 2015), PP 76-81e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[4] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Corpus-Based Concatenative Speech Synthesis System for Marathi" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 20-26e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[5] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[6] Sangramsing Kayte, Dr. Bharti Gawali “Marathi Speech Synthesis: A review” International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6 3708 – 3711

[7] Sangramsing Kayte, Monica Mundada, Jayesh Gujrathi, “ Hidden Markov Model based Speech Synthesis: A Review” International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

[8] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Review of Unit Selection Speech Synthesis International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[9] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Implementation of Marathi Language Speech Databases for Large Dictionary" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 40-45e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[10] Sangramsing Kayte "Transformation of feelings using pitch parameter for Marathi speech" Sangramsing Kayte Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part -4) November 2015, pp.120-124

[11] Sangramsing N.kayte “Marathi Isolated-Word Automatic Speech Recognition System based on Vector Quantization (VQ) approach” 101th Indian Science Congress Jammu University 03th Feb to 07 Feb 2014.

[12] Monica Mundada, Sangramsing Kayte, Dr. Bharti Gawali "Classification of Fluent and Dysfluent Speech Using KNN Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 9, September 2014

[13] Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte” Speech Synthesis System for Marathi Accent using FESTVOX” International Journal of Computer Applications (0975 – 8887) Volume 130 – No.6, November2015

[14] Sangramsing N. Kayte, Dr. Charansing N. Kayte,Dr.Bharti Gawali* "Grapheme-To-Phoneme Tools for the Marathi Speech Synthesis" Sangramsing Kayte et al.Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part -4) November 2015, pp.86-92

[15] Sangramsing N. Kayte,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Rule-based Prosody Calculation for Marathi Text-to-Speech Synthesis" Sangramsing N. Kayte et al. Int. Journal of Engineering Research and Applications ISSN: 2248-

9622, Vol. 5, Issue 11, (Part - 5) November 2015, pp.33-36

[16] Monica Mundada, Sangramsing Kayte "Classification of speech and its related fluency disorders Using KNN" ISSN2231-0096 Volume-4 Number-3 Sept 2014

[17] Monica Mundada, Bharti Gawali, Sangramsing Kayte "Recognition and classification of speech and its related fluency disorders" International Journal of Computer Science and Information Technologies (IJCSIT)

[18] Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte "Screen Readers for Linux and Windows – Concatenation Methods and Unit Selection based Marathi Text to Speech System" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.14, November 2015

[19] [Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[20] Sangramsing N. Kayte,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Implementation of Text To Speech for Marathi Language Using Transcriptions Concept" Sangramsing N. Kayte et al. Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 5) November 2015, pp.33-36

[21] Sangramsing Kayte, Monica Mundada, Santosh Gaikwad, Bharti Gawali "PERFORMANCE EVALUATION OF SPEECH SYNTHESIS TECHNIQUES FOR ENGLISH LANGUAGE " International Congress on Information and Communication Technology 9-10 October, 2015

[22] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte " Performance Calculation of Speech Synthesis Methods for Hindi language IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 13-19e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[23] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Marathi Text-To-Speech Synthesis using Natural Language Processing "IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 63-67e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197

[24] Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte " Performance Evaluation of Speech Synthesis Techniques for Marathi Language " International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

[25] Sangramsing N. Kayte ,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Approach To Build A Marathi Text-To-Speech System Using Concatenative Synthesis Method With The Syllable" Sangramsing Kayte et al.Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part-4) November 2015, pp.93-97

[26] Sangramsing Kayte "Duration for Classification and Regression Tree for Marathi Text-to-Speech Synthesis System" Sangramsing Kayte Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part-4) November2015

[27] Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.Bharti Gawali " Artificially Generated of Concatenative Syllable based Text to Speech Synthesis System for Marathi" OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 44-49e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[28] Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.Bharti Gawali "Automatic Generation of Compound Word Lexicon for Marathi Speech Synthesis" OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver.II (Nov -Dec. 2015), PP 25-30e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[29] Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.Bharti Gawali "Approach of Syllable Based Unit Selection Text-To-Speech Synthesis System for Marathi Using Three Level Fall Back Technique OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 31-35e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197